# LEVERAGING -OMICS BASED APPROACHES TO EXPLORE ENVIRONMENTS: A LOOK AT TWO DOMAINS OF LIFE

By

CHELSEA LOUVOUN MURPHY

Bachelor of Science in Microbiology, Cell and Molecular Biology
Oklahoma State University
Stillwater, Oklahoma
2017

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2022

LEVERAGING -OMICS BASED APPROACHES TO EXPLORE ENVIRONMENTS: A LOOK AT TWO DOMAINS OF LIFE

Dissertation Approved:

Dr. Noha Youssef

Dissertation Advisor

Dr. Mostafa Elshahed

Dr. Wouter Hoff

Dr. Babu Fathepure

Dr. Udaya DeSilva

# ACKNOWLEDGMENTS

First and foremost I would like to thank my advisor Dr. Noha Youssef for all of her help and support throughout these last eight years. From my undergraduate to graduate studies, she has provided immeasurable amounts of guidance and support, and I truly would not be the person and scientist I am today without her. Without her, I may not have fallen in love with microbiology and research, and I definitely would not have accomplished so much. Joining her lab was the best decision I've ever made, and has forever changed my life for the better. Words truly cannot express how profoundly she impacted me. I would also likewise like to thank Dr. Mostafa Elshahed for all of the support these last years as well. Thanks to his contribution, I am not only a better scientist, but also a scientific writer as well.

I would also like to thank my other committee members, Dr. Babu Fathepure, Dr. Wouter Hoff, and Dr. Udaya DeSilva for their support, intriguing discussions, and helpful feedback they've given throughout my journey.

I would also like to acknowledge all the labmates I've had throughout the years who have provided me both personal and professional support. Thanks to Dr. Shelby Calkins, Dr. Ibrahim Farag, and Dr. Radwa Hanafy for providing an example by leading the way. Special thanks to Archana Yadav, who has been a valuable friend throughout my entire doctorate journey. Thanks also to Ryan Hahn, who has been another longtime lab member and friend. Also thanks to Tammy Austin, Carrie Pratt, Adrienne Jones, and Casey Meili for helping to make the lab a fun place of companionship, especially throughout these last few months.

Finally, I would like to thank my family (especially my mom for dealing with all my phone calls), my beloved late rabbit Buttercup and wonderful dog Luca, and my friends for always being there for me. Whenever life got hectic, they've always provided me with the support I needed to make it through, and Luca kept me from being eaten by coyotes during my late night walks through the woods where I did some of my best thinking. Special shoutout to the dog park COVID crew for all those seven hour marathons at the park, giving some levity to an extra stressful time.

I've truly been blessed to be surrounded by so many wonderful people; without them, I don't know if I would have made it this far.

Acknowledgments reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: CHELSEA LOUVOUN MURPHY

Date of Degree: MAY, 2022

Title of Study: LEVERAGING -OMICS BASED APPROACHES TO EXPLORE ENVIRONMENTS: A LOOK AT TWO DOMAINS OF LIFE

Major Field: MICROBIOLOGY, CELL AND MOLECULAR BIOLOGY

Abstract: Rapid advancements in technology have both dramatically lowered the cost of sequencing as well as increased the depth of information gleaned. With such a low barrier of entry, increasing numbers of scientists around the globe are generating unprecedented amounts of data pertaining to the identity and function of the various microbes that impact assorted environments, from within a host to various biomes in nature. Understandably, a pressing challenge in the field centers on how to process and interpret these large quantities of precise information. The leveraging of -omics based techniques used in bioinformatics stands poised to answer this challenge, enabling discoveries that probe not just what a microbe can do, but also perhaps provide a look into their past through an analysis of their genetic potential.

For my overall project, I sought to leverage various -omics based approaches to study how various organisms impact and have been impacted by their respective environments. My work has spread from transcriptomics to proteomics and finally metagenomics, from pure cultures to environmental samples, and from fungi to bacteria. At the surface levels, these works provide a form of functional profile of the studied organisms; however, deeper insights into the potential evolutionary history can also be made. For example, the study on the anaerobic gut fungal phylum Neocallimastigomycota (Chapter I) can provide insights into the clade's intertwined history with the development of herbivory, the study on the obligate plant symbiont *Rhizophagus irregularis* (Chapter II) can provide insight into fungal association with plants, the metagenomic studies of the Binatota (Chapter III) and novel Desulfobacterota classes (Chapter IV) can provide insights into the development and evolution of the Delta Proteobacteria into a diverse clade, and the study of novel Myxococcota classes (Chapter V) can provide clues into the development of predation strategies in bacteria. As a whole, this body of works provides a jumping-off point for future probes into these organisms, as well as potential isolation strategies for the uncultured organisms discussed.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

LIST OF FIGURES

# CHAPTER I

# HORIZONTAL GENE TRANSFER AS AN INDISPENSABLE DRIVER FOR NEOCALLIMASTIGOMYCOTA EVOLUTION INTO A DISTINCT GUT-DWELLING FUNGAL LINEAGE

## 1.1 Abstract

Survival and growth of the anaerobic gut fungi (AGF, Neocallimastigomycota) in the herbivorous gut necessitate the possession of multiple abilities absent in other fungal lineages. We hypothesized that horizontal gene transfer (HGT) was instrumental in forging the evolution of AGF into a phylogenetically distinct gut-dwelling fungal lineage. Patterns of HGT were evaluated in the transcriptomes of 27 AGF strains, 22 of which were isolated and sequenced in this study, and 4 AGF genomes broadly covering the breadth of AGF diversity. We identified 277 distinct incidents of HGT in AGF transcriptomes, with subsequent gene duplication resulting in an HGT frequency of 2-3.5% in AGF genomes. The majority of HGT events were AGF specific (91.7%) and wide (70.8%), indicating their occurrence at early stages of AGF evolution. The acquired genes allowed AGF to expand their substrate utilization range, provided new venues for electron disposal, augmented their biosynthetic capabilities, and facilitated their adaptation to anaerobiosis. The majority of donors were anaerobic fermentative bacteria prevalent in the herbivorous gut. This work strongly indicates that HGT indispensably forged the evolution of AGF as a distinct fungal phylum and provides a unique example of the role of HGT in shaping the evolution of a high rank taxonomic eukaryotic lineage.

## 1.2 Importance

The anaerobic gut fungi (AGF) represent a distinct basal phylum lineage (Neocallimastigomycota) commonly encountered in the rumen and alimentary tracts of herbivores. Survival and growth of anaerobic gut fungi in these anaerobic, eutrophic, and prokaryotes dominated habitats necessitates the acquisition of several traits absent in other fungal lineages. This manuscript assesses the role of horizontal gene transfer as a relatively fast mechanism for trait acquisition by the Neocallimastigomycota post-sequestration in the herbivorous gut. Analysis of twenty-seven transcriptomes that represent the broad Neocallimastigomycota diversity identified 277 distinct HGT events, with subsequent gene duplication resulting in an HGT frequency of 2-3.5% in AGF genomes. These HGT events have allowed AGF to survive in the herbivorous gut by expanding their substrate utilization range, augmenting their biosynthetic pathway, providing new routes for electron disposal by expanding fermentative capacities, and facilitating their adaptation to anaerobiosis. HGT in the AGF is also shown

to be mainly a cross-kingdom affair, with the majority of donors belonging to the bacteria. This work represents a unique example of the role of HGT in shaping the evolution of a high rank taxonomic eukaryotic lineage.

## 1.3 Introduction

Horizontal gene transfer (HGT) is defined as the acquisition, integration, and retention of foreign genetic material into a recipient organism [102]. HGT represents a relatively rapid process for trait acquisition; as opposed to gene creation either from preexisting genes (via duplication, fission, fusion, or exon shuffling) or through *de-novo* gene birth from non-coding sequences [12, 51, 63, 182, 192]. In prokaryotes, the occurrence, patterns, frequency, and impact of HGT on the genomic architecture [284], metabolic abilities [62, 414], physiological preferences [286, 311], and ecological fitness [399] has been widely investigated, and the process is now regarded as a major driver of genome evolution in bacteria and archaea [305, 375]. Although eukaryotes are perceived to evolve principally through modifying existing genetic information, analysis of HGT events in eukaryotic genomes has been eliciting increasing interest and scrutiny. In spite of additional barriers that need to be overcome in eukaryotes, e.g. crossing the nuclear membrane, germline sequestration in sexual multicellular eukaryotes, and epigenetic nucleic acids modifications mechanisms [12, 126], it is now widely accepted that HGT contributes significantly to eukaryotic genome evolution [178, 202]. HGT events have convincingly been documented in multiple phylogenetically disparate eukaryotes ranging from the Excavata [111, 164, 282, 314], SAR supergroup [1, 204, 322, 401], Algae [346], Plants [324], and Opisthokonta [92, 137, 245, 372]. Reported HGT frequency in eukaryotic genomes ranges from a handful of genes, e.g. [249], to up to 9.6% in Bdelloid rotifers [137].

The kingdom Fungi represents a phylogenetically coherent clade that evolved $\approx$ 900-1481 Mya from a unicellular flagellated ancestor [105, 290, 378]. To date, multiple efforts have been reported on the detection and quantification of HGT in fungi. A survey of 60 fungal genomes reported HGT frequencies of 0-0.38% [245], and similar low values were observed in the genomes of five early-diverging pathogenic Microsporidia and Cryptomycota [4]. A recent study has documented the role of HGT in expanding the catabolic capabilities of members of the mycotrophic genus *Trichoderma* by extensive acquisition of plant biomass degradation capacities from plant-associated filamentous Ascomycetes [106]. The osmotrophic lifestyle of fungi [26] has typically been regarded as less conducive to HGT compared to the phagocytic lifestyle of several microeukaryotes with relatively higher HGT frequency [101].

The anaerobic gut fungi (AGF, Neocallimastigomycota) represent a phylogenetically distinct basal fungal lineage. The AGF appear to exhibit a restricted distribution pattern, being encountered in the gut of ruminant and non-ruminant herbivorous [143]. In the herbivorous gut, the life cycle of the AGF (Figure S1) involves the discharge of motile flagellated zoospores from sporangia in response to animal feeding, the chemotaxis and attachment of zoospores to ingested plant material, spore encystment, and the subsequent production of rhizoidal growth that penetrates and digests plant biomass through the production of a wide array of cellulolytic and lignocellulolytic enzymes.

Survival, colonization, and successful propagation of AGF in the herbivorous gut necessitate the acquisition of multiple unique physiological characteristics and metabolic abilities absent in other fungal lineages. These include, but are not limited to, development of a

robust plant biomass degradation machinery, adaptation to anaerobiosis, and exclusive dependence on fermentation for energy generation and recycling of electron carriers [40, 412]. Therefore, we hypothesized that sequestration into the herbivorous gut was conducive to the broad adoption of HGT as a relatively faster adaptive evolutionary strategy for niche adaptation by the AGF (Figure S1). Further, since no part of the AGF life cycle occurs outside the animal host and no reservoir of AGF outside the herbivorous gut has been identified [143], then acquisition would mainly occur from donors that are prevalent in the herbivorous gut (Figure S1). Apart from earlier observations on the putative bacterial origin of a few catabolic genes in two AGF isolates [135, 152], and preliminary BLAST-based queries of a few genomes [148, 412], little is currently known on the patterns, determinants, and frequency of HGT in the Neocallimastigomycota. To address this hypothesis, we systematically evaluated the patterns of HGT acquisition in the transcriptomes of 27 AGF strains and 4 AGF genomes broadly covering the breadth of AGF genus-level diversity. Our results document the high level of HGT in AGF in contrast to HGT paucity across the fungal kingdom. The identity of genes transferred, distribution pattern of events across AGF genera, phylogenetic affiliation of donors, and the expansion of acquired genetic material in AGF genomes highlight the role played by HGT in forging the evolution and diversification of the Neocallimastigomycota as a phylogenetically, metabolically, and ecologically distinct lineage in the fungal kingdom.

## 1.4 Materials and Methods

### 1.4.1 Organisms

Type strains of the Neocallimastigomycota are unavailable through culture collections due to their strict anaerobic and fastidious nature, as well as the frequent occurrence of senescence in AGF strains [165]. As such, obtaining a broad representation of the Neocallimastigomycota necessitated the isolation of representatives of various AGF genera *de novo*. Samples were obtained from the feces, rumen, or digesta of domesticated and wild herbivores around the city of Stillwater, OK and Val Verde County, Texas (Table 1). Samples were immediately transferred to the laboratory and the isolation procedures usually commenced within 24 hours of collection. A second round of isolation was occasionally conducted on samples stored at -$20^0$ C for several weeks (Table 1).

Isolation was performed using a rumen fluid medium reduced by cysteine-sulfide, supplemented with a mixture of kanamycin, penicillin, streptomycin, and chloramphenicol (50 µg/mL, 50 µg/mL, 20 µg/mL, and 50 µg/mL, respectively), and dispensed under a stream of 100% $CO_2$ [150, 412]. All media were prepared according to the Hungate technique [46], as modified by Balch and Wolfe [20]. Cellulose (0.5%), or a mixture of switchgrass (0.5%) and cellobiose (0.5%) were used as carbon sources. Samples were serially diluted and incubated at 39°C for 24-48 h. Colonies were obtained from dilutions showing visible signs of fungal growth using the roll tube technique [175]. Colonies obtained were inoculated into liquid media, and a second round of isolation and colony picking was conducted to ensure culture purity. Microscopic examination of thallus growth pattern, rhizoid morphology, and zoospore flagellation, as well LSU rRNA gene D1-D2 domain amplification and sequencing were employed to determine the genus level affiliation of all isolates [150]. Isolates were main-

tained and routinely sub-cultured on rumen fluid medium supplemented with antibiotics (to guard against accidental bacterial contamination) and stored on agar media as described previously [52, 412].

### 1.4.2 RNA extraction, sequencing, and assembly

Transcriptomic sequencing was conducted for twenty-two AGF strains. Sequencing multiple taxa provides stronger evidence for the occurrence of HGT in a target lineage [323], and allows for the identification of phylum-wide versus genus- and species-specific HGT events. Transcriptomic, rather than genomic, sequencing was chosen for AGF-wide HGT identification efforts since enrichment for polyadenylated (poly(A)) transcripts prior to RNA-seq provides a built-in safeguard against possible prokaryotic contamination, an issue that often plagued eukaryotic genome-based HGT detection efforts [34, 211], as well as to demonstrate that HGT genes identified are transcribed in AGF. Further, sequencing and assembly of a large number of Neocallimastigomycota genomes is challenging due to the extremely high AT content in intergenic regions and the extensive proliferation of microsatellite repeats, often necessitating employing multiple sequencing technologies for successful genomic assembly [148, 412].

Cultures for RNA extraction were grown in rumen fluid medium with cellobiose as the sole carbon source. RNA extraction was conducted on late log/early stationary phase cultures (approximately 48-60 hours post inoculation, depending on strain's growth characteristics) as described previously [82]. Briefly, fungal biomass was obtained by vacuum filtration and grounded with a pestle under liquid nitrogen. RNA was extracted using Epicentre MasterPure Yeast RNA Purification kit (Epicentre, Madison, WI, USA) and stored in RNase-free TE buffer. Transcriptomic sequencing using Illumina HiSeq2500 2X150bp paired end technology was conducted using the services of a commercial provider (Novogene Corporation, Beijing, China).

RNA-Seq reads were assembled by the de novo transcriptomic assembly program Trinity [139] using previously established protocols [145]. All settings were implemented according to the recommended protocol for fungal genomes, with the exception of the absence of the "–jaccard_clip" flag due to the low gene density of anaerobic fungal genomes. The assembly process was conducted on the Oklahoma State University High Performance Computing Cluster as well as the XSEDE HPC Bridges at the Pittsburg Super Computing Center. Quantitative levels for all assembled transcripts were determined using Bowtie2 [224]. The program Kallisto was used for quantification and normalization of the gene expression of the transcriptomes [42]. All final peptide models predicted were annotated using the Trinotate platform with a combination of homology-based search using BLAST+, domain identification using hmmscan and the Pfam 30.0 database 19 [124], and cellular localization with SignalP 4.0 [304]. The twenty-two transcriptomes sequenced in this effort, as well as previously published transcriptomic datasets from *Pecoramyces ruminantium* [412], *Piromyces finnis*, *Piromyces* sp. E2, *Anaeromyces robustus*, and *Neocallimastix californiae* [148] were examined. In each dataset, redundant transcripts were grouped into clusters using CD-HIT-EST with identity parameter of 95% (-c 0.95). The obtained non-redundant transcripts from each analyzed transcriptome were subsequently used for peptide and coding sequence prediction using the TransDecoder with a minimum peptide length of 100 amino

acids (http://transdecoder.github.io). Assessment of transcriptome completeness per strain was conducted using BUSCO [356] using Fungi dataset.

### 1.4.3 HGT identification

A combination of BLAST similarity searches, comparative similarity index (HGT index, $h_U$), and phylogenetic analyses were conducted to identify HGT events in the analyzed transcriptomic datasets (1.1). We define an HGT event as the acquisition of a foreign gene/Pfam by AGF from a single lineage/donor. All predicted peptides were queried against Uniprot databases (downloaded May 2017) each containing both reviewed (Swiss-Prot) and unreviewed (TrEMBL) sequences. The databases encompassed nine different phylogenetic groups; Bacteria, Archaea, Viridiplantae, Opisthokonta-Chaonoflagellida, Opisthokonta-Fungi (without Neocallimastigomycota representatives), Opisthokonta-Metazoa, Opisthokonta- sNucleariidae and Fonticula group, all other Opisthokonta, and all other non-Opisthokonta-non-Viridiplantae Eukaryota. For each peptide sequence, the bit score threshold and HGT index $h_U$ (calculated as the difference between the bit-scores of the best non-fungal and the best Dikarya fungal matches) were determined. Peptide sequences that satisfied the criteria of having a BLASTP bit-score against a non-fungal database that was >100 (i.e. $2^{-100}$ chance of random observation) and an HGT index $h_U$ that was ≥30 were considered HGT candidates and subjected to additional phylogenetic analysis. We chose to work with bit-score rather than the raw scores since the bit-score measures sequence similarity independent of query sequence length and database size. This is essential when comparing hits from databases with different sizes (for example, the Bacteria database contained 83 million sequences while the Choanoflagellida database contained 21 thousand sequences). We chose an $h_U$ value of ≥30 (a difference of bit-score of at least 30 between the best non-fungal hit and the best fungal hit to an AGF sequence) previously suggested and validated [36, 87] as the best tradeoff between sensitivity and specificity. Since the bit-score is a logarithmic value that describes sequence similarity, a bit-score > 30 ensure that the sequence aligned much better to the non-fungal hit than it did to the fungal hit.

The identified HGT candidates were modified by removing all CAZyme-encoding sequences (due to their multi-modular nature, see below) and further clustered into orthologues using OrthoMCL [50]. Orthologues obtained were subjected to detailed phylogenetic analysis to confirm HGT occurrence as well as to determine the potential donor. Each Orthologue was queried against the nr database using web Blastp [54] under two different settings: once against the full nr database and once against the Fungi (taxonomy ID: 4751) excluding the Neocallimastigomycetes (Taxonomy ID: 451455). The first 250 hits obtained using these two Blastp searches with an e-value below $e^{-10}$ were downloaded and combined in one fasta file. To remove redundancies, the downloaded sequences were crudely aligned using the standalone Clustal Omega [354] and the alignements were used to generate phylogenetic trees in FastTree under the LG model [309]. Produced trees were visualized in FigTree and the groups of sequences that clustered together with very short branches were identified. Perl scripts were then used to remove these redundant sequences from the original fasta files (leaving just one representative). The resulting non-redundant fasta files were used for subsequent analysis. AGF and reference sequences were aligned using MAFFT multiple sequence aligner [201], and alignments were masked for sites with >50% alignment gaps

using the Mask Alignment Tool in Geneious 10.2.3 (https://www.geneious.com). Masked alignments were then used in IQ-tree [281] to first predict the best amino acid substitution model (based on the lowest BIC criteria) and to generate maximum likelihood trees under the predicted best model. Both the (-alrt 1000) option for performing the Shimodaira–Hasegawa approximate likelihood ratio test (SH-aLRT), as well as the (-bb 1000) option for ultrafast bootstrap (UFB) [257] were added to the IQ-tree command line. This resulted in the generation of phylogenetic trees with two support values (SH-aLRT and UFB) on each branch. Candidates that showed a nested phylogenetic affiliation that was incongruent to organismal phylogeny with strong SH-aLRT and UFB supports were deemed horizontally transferred. As a final confirmatory step, each tree generated was also reconciled against a species tree (constructed using the large ribosomal subunit L3 protein) using the programs Ranger-DTL [21] and NOTUNG [371] to infer transfer events at the node where AGF taxa clustered with a phylogenetically-incongruent donor.

### 1.4.4 Identification of HGT events in carbohydrate active enzymes (CAZymes) transcripts

In AGF genomes, carbohydrate active enzymes (CAZymes) are often encoded by large multi-module genes with multiple adjacent CAZyme or non-CAZyme domains [148, 412]. A single gene can hence harbor multiple CAZyme pfams of different (fungal or non-fungal) origins [148, 412]. As such, our initial efforts for HGT assessment in CAZyme-encoding transcripts using an entire gene/ transcript strategy yielded inaccurate results since similarity searches only identified pfams with the lowest e-value or highest number of copies, while overlooking additional CAZyme pfams in the transcripts (Figure S2). To circumvent the multi-modular nature of AGF CAZyme transripts, we opted for the identification of CAZyme HGT events on trimmed domains, rather than entire transcript. CAZyme-containing transcripts (Glycoside hydrolases (GHs), Polysaccharide lyases (PLs), and Carbohydrate Esterases (CEs)) were first identified by searching the entire transcriptomic datasets against the dbCAN hidden markov models V5 [410] (downloaded from the dbCAN web server in September 2016) using the command hmmscan in standalone HMMER. For each CAZy family identified, predicted peptides across all transcriptomic datasets were grouped in one fasta file that was then amended with the corresponding Pfam seed sequences (downloaded from the Pfam website (http://pfam.xfam.org/) in March 2017). Sequences were aligned using the standalone Clustal Omega [354] to their corresponding Pfam seeds. Using the Pfam seed sequences as a guide for the start and end of the domain, aligned sequences were then truncated in Jalview [397]. Truncated transcripts with an identified CAZy domain were again compared to the pfam database [123] using hmmscan [308] to ensure correct assignment to CAZy families and accurate domain trimming. These truncated peptide sequences were then analyzed to pinpoint incidents of HGT using the approach described above.

### 1.4.5 Neocallimastigomycota-specific versus non-specific HGT events

To determine whether an identified HGT event (i.e. foreign gene acquisition from a specific donor) is specific to the phylum Neocallimastigomycota; the occurrence of orthologues (30% identity, >100 amino acids alignment) of the identified HGT genes in basal fungi, i.e. mem-

bers of Blastocladiales, Chytridiomycota, Cryptomycota, Microsporidia, Mucormycota, and Zoopagomycota, as well as the putative phylogenetic affiliation of these orthologues, when encountered, were assessed. HGT events were judged to be Neocallimastigomycota-specific if: 1. orthologues were absent in all basal fungal genomes, 2. orthologues were identified in basal fungal genomes, but these orthologues were of clear fungal origin, or 3. orthologues were identified in basal fungal genomes and showed a non-fungal phylogenetic affiliation, but such affiliation was different from that observed in the Neocallimastigomycota. On the other hand, events were judged to be non-specific to the Neocallimastigomycota if phylogenetic analysis of basal fungal orthologues indicated a non-fungal origin with a donor affiliation similar to that observed in the Neocallimastigomycota (1.1).

### 1.4.6 Mapping HGT events to available AGF genomes

HGT events identified in AGF datasets examined (both CAZy and non-CAZy events) were mapped onto currently available AGF genome assemblies [148, 412] (Genbank accession numbers ASRE00000000.1, MCOG00000000.1, MCFG00000000.1, MCFH00000000.1). The duplication and expansion patterns, as well as GC content, and intron distribution were assessed in all identified genes. Averages were compared to AGF genome average using Student t-test to identify possible deviations in such characteristics as often observed with HGT genes [363]. To avoid any bias the differences in the number of genes compared might have on the results, we also compared the GC content, codon usage, and intron distribution averages for the identified genes to a subset of an equal number of randomly chosen genes from AGF genomes. We used the MEME Suite's fasta-subsample function (http://meme-suite.org/doc/fasta-subsample.html) to randomly select an equal number of genes from the AGF genomes.

### 1.4.7 Validation of HGT-identification pipeline using previously published datasets

As a control, the frequency of HGT occurrence in the genomes of a filamentous ascomycete (*Colletotrichum graminicola,* GenBank Assembly accession number GCA_000149035.1), and a microsporidian (*Encephalitozoon hellem,* GenBank Assembly accession number GCA_000277815.3) were determined using our pipeline (Table S1); and the results were compared to previously published results [4, 187].

### 1.4.8 Guarding against false positive HGT events due to contamination

Multiple safeguards were taken to ensure that the frequency and incidence of HGT reported here are not due to bacterial contamination of AGF transcripts. These included: 1. Application of antibiotics in all culturing procedures as described above. 2. Utilization of transcriptomes rather than genomes selects for eukaryotic polyadenylated (poly(A)) transcripts prior to RNA-seq as a built-in safeguard against possible prokaryotic contamination. 3. Mapping HGT transcripts identified to genomes generated in prior studies and confirming the occurrence of introns in the majority of HGT genes identified. 4. Applying a threshold where only transcripts identified in >50% of transcriptomic assemblies from a specific genus are included and 5. The exclusion of HGT events showing suspiciously high (>90%) sequence identity to donor sequences.

In addition, recent studies have demonstrated that GenBank-deposited reference genomes [34] and transcriptomes [368] of multicellular organisms are often plagued by prokaryotic contamination. The occurrence of prokaryotic contamination in reference donors' genomes/ transcriptomes could lead to false positive HGT identification, or incorrect HGT assignments. To guard against any false positive HGT event identification due to possible contamination in reference datasets, sequence data from potential donor reference organisms were queried using blast, and their congruence with organismal phylogeny was considered a prerequisite for inclusion of an HGT event.

### 1.4.9 Data accession

Sequences of individual transcripts identified as horizontally transferred are deposited in GenBank under the accession number MH043627-MH043936, and MH044722-MH044724. The whole transcriptome shotgun sequences were deposited in GenBank under the BioProject PRJNA489922, and Biosample accession numbers SAMN09994575- SAMN09994596. Transcriptomic assemblies were deposited in the SRA under project accession number SRP161496. Trees of HGT events discussed in the results and discussion sections are presented in the supplementary document (S5-S45).

## 1.5 Results

### 1.5.1 Isolates

The transcriptomes of 22 different isolates were sequenced. These isolates belonged to six out of the nine currently described AGF genera: *Anaeromyces* (n=5), *Caecomyces* (n=2), *Neocallimastix* (n=2), *Orpinomyces* (n=3), *Pecoramyces* (n=4), *Piromyces* (n=4), as well as the recently proposed genus *Feramyces* (n=2) [151] (Table 1, Supplementary Fig. 3). Out of the three AGF genera not included in this analysis, two are currently represented by a single strain that was either lost (genus *Oontomyces* [90]), or appears to exhibit an extremely limited geographic and animal host distribution (genus *Buwchfawromyces* [53]). The third unrepresented genus (*Cyllamyces*) has recently been suggested to be phylogenetically synonymous with *Caecomyces* [393]. As such, the current collection is a broad representation of currently described AGF genera.

### 1.5.2 Sequencing

Transcriptomic sequencing yielded 15.2-110.8 million reads (average, 40.87) that were assembled into 31,021-178,809 total transcripts, 17,539-132,141 distinct transcripts (clustering at 95%), and 16,500-70,061 predicted peptides (average 31,611) (Table S2). Assessment of transcriptome completion using BUSCO [356] yielded high values (82.76-97.24%) for all assemblies (Table S1). For strains with a sequenced genome, genome coverage (percentage of genes in a strain's genome for which a transcript was identified) ranged between 70.9-91.4% (Table S2).

### 1.5.3 HGT events

A total of 12,786 orthologues with a non-fungal bit score > 100, and an HGT index > 30 were identified. After removing orthologues occurring only in a single strain or in less than 50% of isolates belonging to the same genus, 2147 events were further evaluated. Phylogenetic analysis could not confirm the HGT nature (e.g. single long branch that could either be attributed to HGT or gene loss in all other fungi, unstable phylogeny, and/or low bootstrap) of 1863 orthologues and so were subsequently removed. Of the remaining 286 orthologues, 8 had suspiciously high (>90%) first hit amino acid identity. Although the relatively recent divergence and/or acquisition time could explain this high level of similarity, we opted to remove these orthologues as a safeguard against possible bacterial contamination of the transcriptomes. Of the remaining 278 orthologues, one was not inferred as horizontally transferred by the gene-species trees reconciliation softwares used. Ultimately, a total of 277 distinct HGT events that satisfied the criteria described above for HGT were identified (Table S3). The average number of events per genus was 220±12.6 and ranged between 206 in the genus *Orpinomyces* to 237 in the genus *Pecoramyces* pantranscriptomes (1.2A). The majority of HGT acquisition events identified (254, 91.7%) appear to be Neocallimastigomycota-specific, i.e. identified only in genomes belonging to the Neocallimastigomycota, but not in other basal fungal genomes (Table S4), strongly suggesting that such acquisitions occurred post, or concurrent with, the evolution of Neocallimastigomycota as a distinct fungal lineage. As well, the majority of these identified genes were Neocallimastigomycota-wide, being identified in strains belonging to at least six out of the seven examined genera (196 events, 70.76%), suggesting the acquisition of such genes prior to genus level diversification within the Neocallimastigomycota. Only 30 events (10.83%) were genus-specific, with the remainder (51 events, 18.4%) being identified in the transcriptomes of 3-5 genera (Table S4, Figure S4, and 1.2b).

The absolute majority (89.2%) of events were successfully mapped to at least one of the four AGF genomes (Table S5), with a fraction (7/30) of the unmapped transcripts being specific to a genus with no genome representative (*Feramyces*, *Caecomyces*). Compared to a random subset of 277 genes in each of the sequenced genomes, horizontally transferred genes in AGF genomes exhibited significantly (P<0.0001) fewer introns (1.1±0.31 vs 3.32±0.83), as well as higher GC content (31±4.5 vs 27.7±5.5) (Table S5). Further, HGT genes/pfams often displayed high levels of gene/ pfam duplication and expansion within the genome (Table S5), resulting in an HGT frequency of 2.03% in *Pecoramyces ruminantium* (331 HGT genes out of 16,347 total genes), 2.91% in *Piromyces finnis* (334 HGT genes out of 11,477 total genes), 3.21% in *Anaeromyces robustus* (415 HGT genes out of 12,939 total genes), and 3.46% in *Neocallimastix californiae* (724 HGT genes out of 20,939 total genes).

### 1.5.4 Donors

A bacterial origin was identified for the majority of HGT events (85.9%), with four bacterial phyla (Firmicutes, Proteobacteria, Bacteroidetes, and Spirochaetes) identified as donors for 169 events (61% of total, 71% of bacterial events) (1.3A). Specifically, the contribution of members of the Firmicutes (119 events) was paramount, the majority of which were most closely affiliated with members of the order Clostridiales (93 events). In addition, minor

contributions from a wide range of bacterial phyla were also identified (1.3A). The majority of the putative donor taxa are strict/ facultative anaerobes, and many of which are also known to be major inhabitants of the herbivorous gut and often possess polysaccharide-degradation capabilities [157, 370]. Archaeal contributions to HGT were extremely rare (5 events). On the other hand, multiple (30) events with eukaryotic donors were identified. In few instances, a clear non-fungal origin was identified for a specific event, but the precise inference of the donor based on phylogenetic analysis was not feasible (Table S4).

### 1.5.5 Metabolic characterization

Functional annotation of HGT genes/pfams indicated that the majority (63.9%) of events encode metabolic functions such as extracellular polysaccharide degradation and central metabolic processes. Bacterial donors were slightly overrepresented in metabolic HGT events (87.5% of the metabolism-related events, compared to 85.9% of the total events). Genes involved in cellular processes and signaling represent the second most represented HGT events (11.19%), while genes involved in information storage and processing only made up 4.69% of the HGT events identified (Figs 3b-e). Below we present a detailed description of the putative abilities and functions enabled by HGT transfer events.

#### Central catabolic abilities

Multiple HGT events encoding various central catabolic processes were identified in AGF transcriptomes and successfully mapped to the genomes (Fig. 4, Table S4, Figs S5-S16). A group of events appears to encode enzymes that allow AGF to channel specific substrates into central metabolic pathways. For example, genes encoding enzymes of the Leloir pathway for galactose conversion to glucose-1-phosphate (galactose-1-epimerase, galactokinase (Fig. 5A), and galactose-1-phosphate uridylyltransferase) were identified, in addition to genes encoding ribokinase, as well as xylose isomerase and xylulokinase for ribose and xylose channeling into the pentose phosphate pathway. As well, genes encoding deoxyribose-phosphate aldolase (DeoC) enabling the utilization of purines as carbon and energy sources were also horizontally acquired in AGF. Further, several of the glycolysis/gluconeogenesis genes, e.g. phosphoenolpyruvate synthase, as well as phosphoglycerate mutase were also of bacterial origin. Fungal homologs of these glycolysis/gluconeogenesis genes were not identified in the AGF transcriptomes and genomes, suggesting the occurrence of xenologous replacement HGT events.

In addition to broadening substrate range, HGT acquisitions provided additional venues for recycling reduced electron carriers via new fermentative pathways in this strictly anaerobic and fermentative lineage. The production of ethanol, D-lactate, and hydrogen appears to be enabled by HGT (Fig. 4). The acquisition of several aldehyde/alcohol dehydrogenases, and of D-Lactate dehydrogenase for ethanol and lactate production from pyruvate was identified. Although these two enzymes are encoded in other fungi as part of their fermentative capacity (e.g. *Saccharomyces* and *Schizosaccharomyces*), no homologs of these fungal genes were identified in AGF pantranscriptomes. Hydrogen production in AGF, as well as in many anaerobic eukaryotes with mitochondria-related organelles (e.g. hydrogenosomes and mitosomes), involves pyruvate decarboxylation to acetyl CoA, followed by the use of electrons

generated for hydrogen formation via an anaerobic Fe-Fe hydrogenase [146, 273, 413]. In AGF, while enzymes for pyruvate decarboxylation to acetyl CoA (pyruvate-formate lyase) and the subsequent production of acetate in the hydrogenosome (via acetyl-CoA:succinyl transferase) appear to be of fungal origin, the Fe-Fe hydrogenase and its entire maturation machinery (HydEFG) seem to be horizontally transferred being phylogenetically affiliated with similar enzymes in Thermotogae, Clostridiales, and the anaerobic jakobid excavate, *Stygiella incarcerate* (Fig. 5B). It has recently been suggested that *Stygiella* acquired the Fe-Fe hydrogenase and its maturation machinery from bacterial donors including Thermotogae, Firmicutes, and Spirochaetes [229], suggesting either *a* single early acquisition event in eukaryotes, or alternatively independent events for the same group of genes have occurred in different eukaryotes. With the exception of the Fe-Fe hydrogenase and its maturation machinery, no other hydrogenosomally-destined proteins (see list in reference [412]) were identified as horizontally transferred in this study. These results collectively suggest that HGT did not play a role in the evolution of hydrogenosomes in AGF; and reinforces the proposed mitochondrial origin of hydrogenosomes through reductive evolution [146].

### Anabolic capabilities

Multiple anabolic genes that expanded AGF biosynthetic capacities appear to be horizontally transferred (Fig. S17-S30). These include several amino acid biosynthesis genes e.g. cysteine biosynthesis from serine; glycine and threonine interconversion; and asparagine synthesis from aspartate. In addition, horizontal gene transfer allowed AGF to de-novo synthesize NAD via the bacterial pathway (starting from aspartate via L-aspartate oxidase (NadB; Fig. 5C) and quinolinate synthase (NadA) rather than the five-enzymes fungal pathway starting from tryptophan [236]). HGT also allowed AGF to salvage thiamine via the acquisition of phosphomethylpyrimidine kinase. Additionally, several genes encoding enzymes in purine and pyrimidine biosynthesis were horizontally transferred (Fig. 4). Finally, horizontal gene transfer allowed AGF to synthesize phosphatidyl-serine from CDP-diacylglycerol, and to convert phosphatidyl-ethanolamine to phosphatidyl-choline.

### Adaptation to the host environment

Horizontal gene transfer also appears to have provided means of guarding against toxic levels of compounds known to occur in the host animal gut (Fig. S31-S37). For example, methylglyoxal, a reactive electrophilic species [227], is inevitably produced by ruminal bacteria from dihydroxyacetone phosphate when experiencing growth conditions with excess sugar and limiting nitrogen [340]. Genes encoding enzymes mediating methylglyoxal conversion to D-lactate (glyoxalase I and glyoxalase II-encoding genes) appear to be acquired via HGT in AGF. Further, HGT allowed several means of adaptation to anaerobiosis. These include: 1) acquisition of the oxygen-sensitive ribonucleoside-triphosphate reductase class III (Fig. 5D) that is known to only function during anaerobiosis to convert ribonucleotides to deoxyribonucleotides [190], 2) acquisition of squalene-hopene cyclase, which catalyzes the cyclization of squalene into hopene, an essential step in biosynthesis of the cell membrane steroid tetrahymanol that replaced the molecular $O_2$-requiring ergosterol in the cell membranes of AGF, 3) acquisition of several enzymes in the oxidative stress machinery including Fe/Mn superoxide

dismutase, glutathione peroxidase, rubredoxin/rubrerythrin, and alkylhydroperoxidase.

In addition to anaerobiosis, multiple horizontally transferred general stress and repair enzymes were identified (Fig. S38-S45). HGT-acquired genes encoding 2-phosphoglycolate phosphatase, known to metabolize the 2-phosphoglycolate produced in the repair of DNA lesions induced by oxidative stress [303] to glycolate, were identified in all AGF transcriptomes studied (Fig. 4, Table S4). Surprisingly, two genes encoding antibiotic resistance enzymes, chloramphenicol acetyltransferase and aminoglycoside phosphotransferase, were identified in all AGF transcriptomes, presumably to improve its fitness in the eutrophic rumen habitat that harbors antibiotic-producing prokaryotes (Table S4). While unusual for eukaryotes to express antibiotic resistance genes, basal fungi such as *Allomyces, Batrachochytrium,* and *Blastocladiella* were shown to be susceptible to chloramphenicol and streptomycin [29, 333]. Other horizontally transferred repair enzymes include DNA-3-methyladenine glycosylase I, methylated-DNA--protein-cysteine methyltransferase, galactoside and maltose O-acetyltransferase, and methionine-R-sulfoxide reductase (Table S4).

**HGT transfer in AGF carbohydrate active enzymes machinery**

Within the analyzed AGF transcriptomes, CAZymes belonging to 39 glycoside hydrolase (GHs), 5 polysaccharide lyase (PLs), and 10 carbohydrate esterase (CEs) families were identified (Fig. 6). The composition of the CAZymes of various AGF strains examined were broadly similar, with the following ten notable exceptions: Presence of GH24 and GH78 transcripts only in *Anaeromyces* and *Orpinomyces,* the presence of GH28 transcripts only in *Pecoramyces*, *Neocallimastix*, and *Orpinomyces*, the presence of GH30 transcripts only in *Anaeromyces*, and *Neocallimastix*, the presence of GH36 and GH95 transcripts only in *Anaeromyces*, *Neocallimastix*, and *Orpinomyces,* the presence of GH97 transcripts only in *Neocallimastix*, and *Feramyces*, the presence of GH108 transcripts only in *Neocallimastix*, and *Piromyces*, and the presence of GH37 predominantly in *Neocallimastix*, GH57 transcripts predominantly in *Orpinomyces*, GH76 transcripts predominantly in *Feramyces*, and CE7 transcripts predominantly in *Anaeromyces* (Fig. 6).

HGT appears to be rampant in the AGF pan-CAZyome: A total of 72 events (26% of total HGT events) were identified, with 40.3% occurring in at least 6 of the 7 AGF genera examined (Fig. 6, Table S4). In 48.7% of GH families, 50% of CE families, and 40% of PL families, a single event (i.e. attributed to one donor) was observed (Fig. 6, Table S4).

Duplication of these events in AGF genomes was notable, with 132, 310, 156, and 130 copies of HGT CAZyme pfams identified in *Anaeromyces*, *Neocallimastix*, *Piromyces* and *Pecoramyces* genomes, representing 33.59%, 36.77%, 40.41%, and 24.62% of the overall organismal CAZyme machinery (Table S5). The contribution of Viridiplantae, Fibrobacteres, and Gamma-Proteobacteria was either exclusive to CAZyme-related HGT events or significantly higher in CAZyme, compared to other events (1.3A).

Transcripts acquired by HGT represented >50% of transcripts in anywhere between 13 (*Caecomyces*) to 20 (*Anaeromyces*) GH families; 3 (*Caecomyces*) to 5 (*Anaeromyces*, *Neocallimastix*, *Orpinomyces*, and *Feramyces*) CE families; and 2 (*Caecomyces* and *Feramyces*) to 3 (*Anaeromyces*, *Pecoramyces*, *Piromyces*, *Neocallimastix*, and *Orpinomyces*) PL families (Fig. 6). It is important to note that in all these families, multiple transcripts appeared to be of bacterial origin based on BLAST similarity search but did not meet the strict crite-

ria implemented for HGT determination in this study. As such, the contribution of HGT transcripts to overall transcripts in these families is probably an underestimate. Only GH9, GH20, GH37, GH45, and PL3 families appear to lack any detectable HGT events. A PCA biplot comparing CAZyomes in AGF genomes to other basal fungal lineages strongly suggests that the acquisition and expansion of many of these foreign genes play an important role in shaping the lignocellulolytic machinery of AGF (Fig. 7). The majority of CAZyme families defining AGF CAZyome were predominantly of non-fungal origin (Fig. 7). This pattern clearly attests to the value of HGT in shaping AGF CAZyome via acquisition and extensive duplication of acquired gene families.

Collectively, HGT had a profound impact on AGF plant biomass degradation capabilities, as recently proposed [108]. The AGF CAZyome encodes enzymes putatively mediating the degradation of twelve different polysaccharides (Fig. S46). In all instances, GH and PL families with >50% horizontally transferred transcripts contributed to backbone cleavage of these polymers; although in many polymers, e.g. cellulose, glucoarabinoxylan, and rhamnogalactouronan, multiple different GHs can contribute to backbone cleavage. Similarly, GH, CE, and PL families with >50% horizontally transferred transcripts contributed to 10 out of 13 side-chain-cleaving activities, and 3 out of 5 oligomer-to-monomer breakdown activities (Fig. S46).

### 1.6 Discussion

Here, we present a systematic analysis of HGT patterns in 27 transcriptomes and 4 genomes belonging to the Neocallimastigomycota. Our analysis identified 277 events, representing 2-3.46% of genes in examined AGF genomes. Further, we consider these values to be conservative estimates due to the highly stringent criteria and employed. Only events with $h_U$ of >30 were considered, and all putative events were further subjected to manual inspection, phylogenetic tree construction, and gene-species tree reconciliation analysis to confirm incongruence with organismal evolution and bootstrap-supported affiliation to donor lineages. Further, events identified in less than 50% of strains in a specific genus were excluded, and parametric gene composition approaches were implemented in conjunction with sequence-based analysis.

Multiple factors could be postulated to account for the observed high HGT frequency in AGF. The sequestration of AGF into the anaerobic, prokaryotes-dominated herbivorous gut necessitated the implementation of the relatively faster adaptive mechanisms for survival in this new environment, as opposed to the slower strategies of neofunctionalization and gene birth. Indeed, niche adaptation and habitat diversification events are widely considered important drivers for HGT in eukaryotes [94, 202, 322, 346] [106]. Further, AGF are constantly exposed to a rich milieu of cells and degraded DNA in the herbivorous gut. Such close physical proximity between donors/ extracellular DNA and recipients is also known to greatly facilitate HGT [25, 263, 352]. Finally, AGF release asexual motile free zoospores into the herbivorous gut as part of their life cycle [143]. According to the weak-link model [170], these weakly protected and exposed structures provide excellent entry point of foreign DNA to eukaryotic genomes. It is important to note that AGF zoospores also appear to be naturally competent, capable of readily uptaking nucleic acids from their surrounding environment [52].

The anaerobic gut fungi have a notoriously low GC content, ranging between 13-20%. It has previously been postulated that this low GC content is due to genetic drift [412] triggered by the low effective population sizes, bottlenecks in vertical transmission, and the asexual life style of anaerobic fungi. As such, the low GC content is an additional consequence of AGF sequestration in the herbivorous gut. Whether the low GC content in AGF played a role in facilitating HGT is currently unclear. It is worth mentioning, however, that the majority of AGF donors identified in this study are members of the bacterial order Clostridiales, many of which have relatively low GC content genomes.

The distribution of HGT events across various AGF taxa (1.2), identities of HGT donors (1.3), and abilities imparted (Figs. 4-5) could offer important clues regarding the timing and impact of HGT on Neocallimastigomycota evolution. The majority of events (70.76%) were Neocallimastigomycota-wide and were mostly acquired from lineages known to inhabit the herbivorous gut, e.g. Firmicutes, Proteobacteria, Bacteroidetes, and Spirochaetes (Figs. 2-3). This pattern strongly suggests that such acquisitions occurred post (or concurrent with) AGF sequestration into the herbivorous gut, but prior to AGF genus level diversification. Many of the functions encoded by these events represented novel functional acquisitions that impart new abilities, e.g. galactose metabolism, methyl glyoxal detoxification, pyruvate fermentation to d-lactate and ethanol, and chloramphenicol resistance (1.3). Others represented acquisition of novel genes or pfams augmenting existing capabilities within the AGF genomes, e.g. acquisition of GH5 cellulases to augment the fungal GH45, acquisition of additional GH1 and GH3 beta gluco- and galactosidases to augment similar enzymes of apparent fungal origin in AGF genomes (Fig. 6-7, Fig. S46). Novel functional acquisition events enabled AGF to survive and colonize the herbivorous gut by: 1. Expanding substrate-degradation capabilities (Fig. 5a, 6, 7, S5-S17, Table S4), hence improving fitness by maximizing carbon and energy acquisition from available plant substrates, 2. Providing additional venues for electron disposal via lactate, ethanol, and hydrogen production, and 3. Enabling adaptation to anaerobiosis (Fig. 4, S32-S38, Table S4).

A smaller number of observed events (n=30) were genus-specific (1.2, Table S4). This group was characterized by being significantly enriched in CAZymes (56.7% of genus-specific horizontally transferred events have a predicted CAZyme function, as opposed to 26% in the overall HGT dataset), and being almost exclusively acquired from donors that are known to inhabit the herbivorous gut [86] (25 out of the 30 events were acquired from the orders Clostridiales, Bacillales, and Lactobacillales within Firmicutes, Burkholderiales within the Beta-Proteobacteria, Flavobacteriales and Bacteroidales within Bacteroidetes, and the Spirochaetes, Actinobacteria, and Lentisphaerae), or from Viridiplantae (4 out of the 30 events). Such pattern suggests the occurrence of these events relatively recently in the herbivorous gut post AGF genus level diversification. A recent study also highlighted the role of HGT in complementing the CAZyme machinery of *Piromyces* sp. strain E2 [108]. We reason that the lower frequency of such events is a reflection of the relaxed pressure for acquisition and retention of foreign genes at this stage of AGF evolution.

Gene acquisition by HGT necessitates physical contact between donor and recipient organisms. Many of the HGT acquired traits by AGF are acquired from prokaryotes that are prevalent in the herbivorous gut microbiota (1.3). However, since many of these traits are absolutely necessary for survival in the gut, the establishment of AGF ancestors in this seemingly inhospitable habitat is, theoretically, unfeasible. This dilemma is common to all

HGT processes enabling niche adaptation and habitat diversification [115]. We put forth two evolutionary scenarios that could explain this dilemma not only for AGF, but also for other gut-dwelling anaerobic microeukaryotes, e.g. *Giardia, Blastocystis*, and *Entamoeba*, where HGT was shown to play a vital role in enabling survival in anaerobic conditions [11, 115, 141]. The first is a coevolution scenario in which the progressive evolution of the mammalian gut from a short and predominantly aerobic structure characteristic of carnivores/insectivores to the longer, more complex, and compartmentalized structure encountered in herbivores was associated with a parallel progressive and stepwise acquisition of genes required for plant polymers metabolism and anaerobiosis by AGF ancestors, hence assuring its survival and establishment in the current herbivorous gut. The second possibility is that AGF ancestors were indeed acquired into a complex and anaerobic herbivorous gut, but initially represented an extremely minor component of the gut microbiome and survived in locations with relatively higher oxygen concentration in the alimentary tract e.g. mouth, saliva, esophagus or in micro-niches in the rumen where transient oxygen exposure occurs. Subsequently, HGT acquisition has enabled the expansion of their niche, improved their competitiveness and their relative abundance in the herbivorous gut to the current levels. In conclusion, our survey of HGT in AGF acquisition demonstrates that the process is absolutely crucial for the survival and growth of AGF in its unique habitat. This is not only reflected in the large number of events, massive duplication of acquired genes, and overall high HGT frequency observed in AGF genomes, but also in the nature of abilities imparted by the process. HGT events not only facilitated AGF adaptation to anaerobiosis, but also allowed them to drastically improve their polysaccharide degradation capacities, provide new venues for electron disposal via fermentation, and acquire new biosynthetic abilities. As such, we reason that the process should not merely be regarded as a conduit for supplemental acquisition of few additional beneficial traits. Rather, we posit that HGT enabled AGF to forge a new evolutionary trajectory, resulting in Neocallimastigomycota sequestration, evolution as a distinct fungal lineage in the fungal tree of life, and subsequent genus and species level diversification. This provides an excellent example of the role of HGT in forging the formation of high rank taxonomic lineages during eukaryotic evolution.

## 1.7 Acknowledgements

This work has been funded by the NSF-DEB Grant numbers 1557102 to N.Y. and M.E. and 1557110 to J.E.S.

This work is published and available in full online (DOI: 10.1128/AEM.00988-19).

## 1.8 Figures & Tables

Supplementary Figures and Tables can be viewed online at:
https://journals.asm.org/doi/10.1128/AEM.00988-19

| Genus | Species | Strain | Host | Isolation source | Location | LSU Genbank accession number | Reference |
|---|---|---|---|---|---|---|---|
| *Anaeromyces* | *contortus* | C3G | Cow (*Bos taurus*) | Feces | Stillwater, OK | MF121936 | This study |
| *Anaeromyces* | *contortus* | C3J | Cow (*Bos taurus*) | Feces | Stillwater, OK | MF121942 | This study |
| *Anaeromyces* | *contortus* | G3G | Goat (*Capra aegagrus hircus*) | Feces | Stillwater, OK | MF121935 | This study |
| *Anaeromyces* | *contortus* | Na | Cow (*Bos taurus*) | Feces | Stillwater, OK | MF121943 | This study |
| *Anaeromyces* | *contortus* | O2 | Cow (*Bos taurus*) | Feces | Stillwater, OK | MF121931 | This study |
| *Anaeromyces* | *robustus* | S4 | Sheep (*Ovis aries*) | Feces | Santa Barbara, CA | NA* | (45) |
| | | | | | | | |
| *Caecomyces* | sp. | Iso3 | Cow (*Bos taurus*) | Feces | Stillwater, OK | MG992499 | This study |
| *Caecomyces* | sp. | Brit4 | Cow (*Bos taurus*) | Rumen | Stillwater, OK | MG992500 | This study |
| | | | | | | | |
| *Feramyces* | *austinii* | F2c | Aoudad sheep (*Ammotragus lervia*) | Feces | Stillwater, OK | MG605675 | This study |
| *Feramyces* | *austinii* | F3a | Aoudad sheep (*Ammotragus lervia*) | Feces | Stillwater, OK | MG584226 | This study |
| | | | | | | | |
| *Neocallimastix* | *californiae* | G1 | Goat (*Capra aegagrus hircus*) | Feces | Santa Barbara, CA | Genomic sequence** | (45) |
| *Neocallimastix* | cf. *cameroonii* | G3 | Sheep (*Ovis aries*) | Feces | Stillwater, OK | MG992493 | This study |
| *Neocallimastix* | cf. *frontalis* | Hef5 | Cow (*Bos taurus*) | Feces | Stillwater, OK | MG992494 | This study |
| | | | | | | | |
| *Orpinomyces* | cf. *joyonii* | D3A | Cow (*Bos taurus*) | Digesta | Stillwater, OK | MG992487 | This study |
| *Orpinomyces* | cf. *joyonii* | D3B | Cow (*Bos taurus*) | Digesta | Stillwater, OK | MG992488 | This study |
| *Orpinomyces* | cf. *joyonii* | D4C | Cow (*Bos taurus*) | Digesta | Stillwater, OK | MG992489 | This study |
| | | | | | | | |
| *Pecoramyces* | *ruminantium* | C1A | Cow (*Bos taurus*) | Feces | Stillwater, OK | JN939127 | (42, 55) |
| *Pecoramyces* | *ruminantium* | S4B | Sheep (*Ovis aries*) | Feces | Stillwater, OK | KX961618 | This study |
| *Pecoramyces* | *ruminantium* | FS3C | Cow (*Bos taurus*) | Rumen | Stillwater, OK | MG992492 | This study |
| *Pecoramyces* | *ruminantium* | FX4B | Cow (*Bos taurus*) | Rumen | Stillwater, OK | MG992491 | This study |
| *Pecoramyces* | *ruminantium* | YC3 | Cow (*Bos taurus*) | Rumen | Stillwater, OK | MG992490 | This study |
| | | | | | | | |
| *Piromyces* | *finnis* | finn | Horse (*Equus caballus*) | Feces | Santa Barbara, CA | Genomic sequence** | (45) |
| *Piromyces* | sp. | A1 | Sheep (*Ovis aries*) | Feces | Stillwater, OK | MG992496 | This study |
| *Piromyces* | sp. | A2 | Sheep (*Ovis aries*) | Feces | Stillwater, OK | MG992495 | This study |
| *Piromyces* | sp. | B4 | Cow (*Bos taurus*) | Feces | Stillwater, OK | MG992497 | This study |
| *Piromyces* | sp. | B5 | Cow (*Bos taurus*) | Feces | Stillwater, OK | MG992498 | This study |
| *Piromyces* | sp. | E2 | Indian Elephant (*Elephas maximus*) | Feces | London, UK | NA | (45, 106) |
| | | | | | | | |

*NA: Not available

** LSU sequence was extracted from the genomic assembly. No LSU accession number was available.

Table 1.1: Neocallimastigomycota strains analyzed in this study.

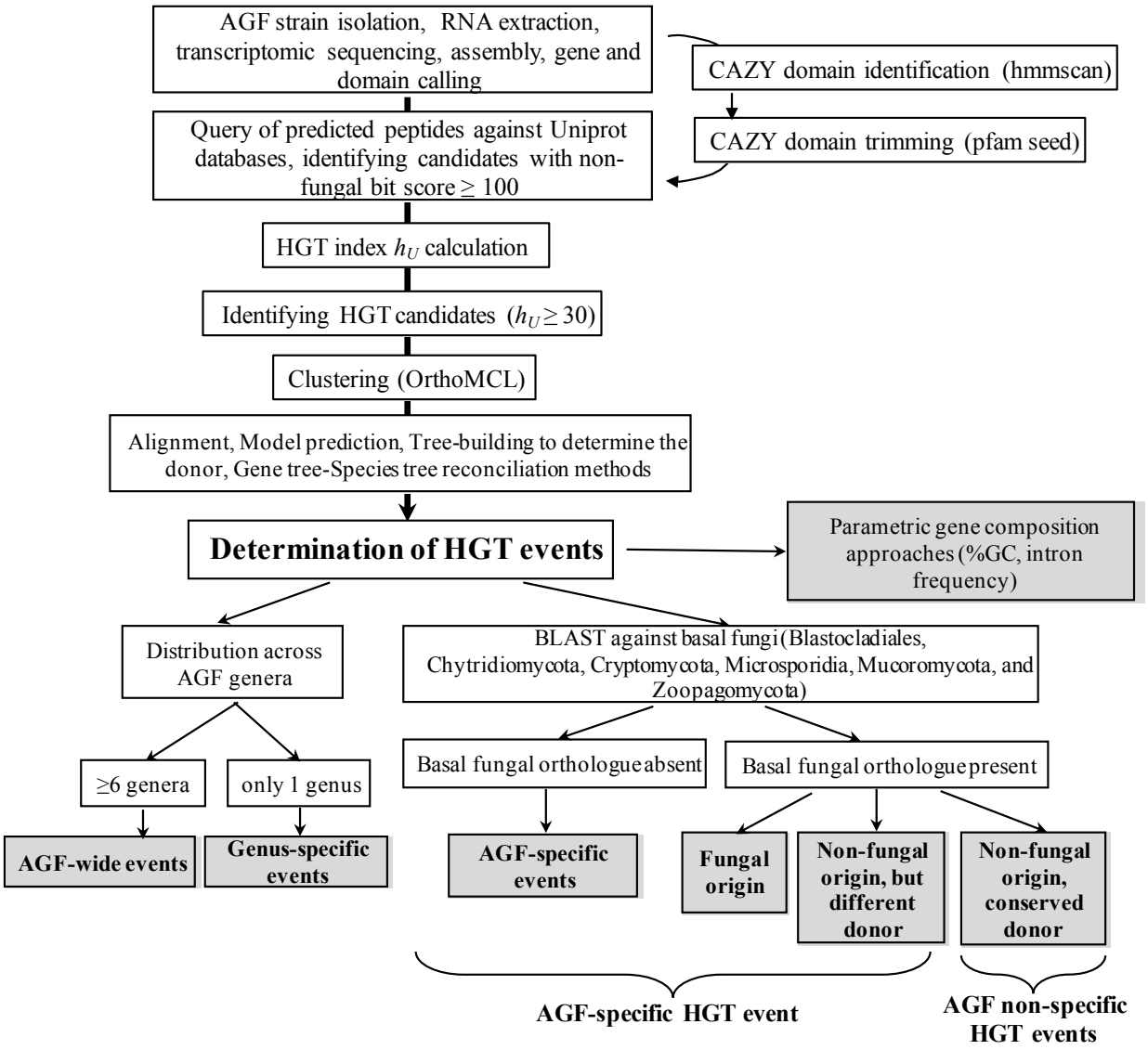



Figure 1.1: Workflow diagram describing the procedure employed for identification HGT events in Neocallimastigomycota datasets analyzed in this study.

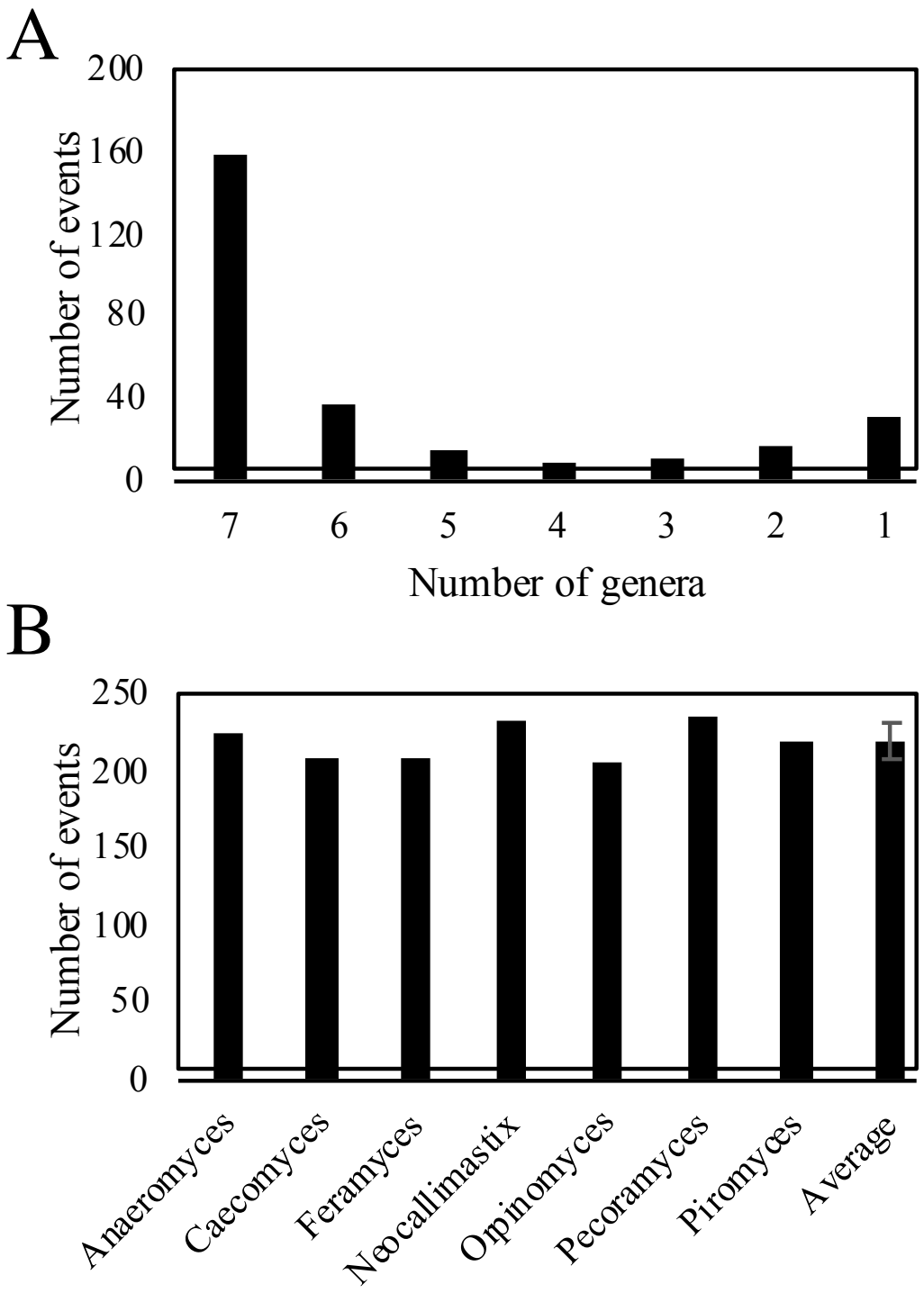



Figure 1.2: **(A)** Distribution pattern of HGT events in AGF transcriptomes demonstrating that the majority of events were Neocallimastigomycota-wide i.e. identified in all seven AGF genera examined. **(B)** Total Number of HGT events identified per AGF genus.

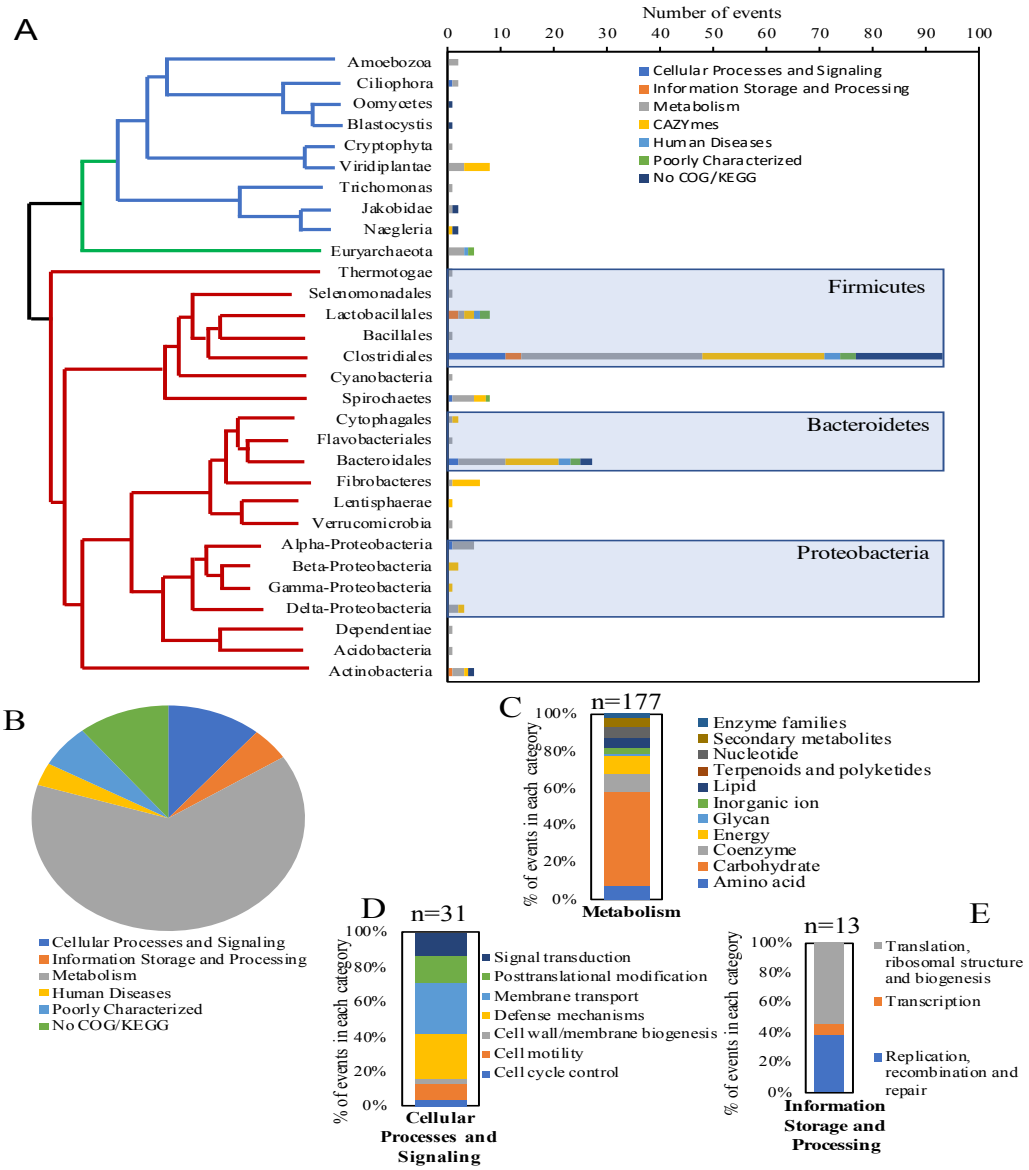



Figure 1.3: Identity of HGT donors and their contribution to the various functional classes. The X-axis shows the absolute number of events belonging to each of the functional classes shown in the legend. The tree is intended to show the relationship between the donors' taxa and is not drawn to scale. Bacterial donors are shown with red branches depicting the phylum-level, with the exception of Firmicutes and Bacteroidetes donors, where the order-level is shown, and Proteobacteria, where the class-level is shown. Archaeal donors are shown with green branches and all belonged to the Methanobacteriales order of Euryarchaeota. Eukaryotic donors are shown with blue branches. Only the 230 events from a definitive-taxon donor are shown in the figure. The other 53 events were clearly nested within a non-fungal clade, but a definitive donor taxon could not be ascertained. Functional classification of the HGT events, determined by searching the Conserved Domain server [246] against the COG database are shown in **B**. For events with no COG classification, a search against the KEGG orthology database [198] was performed. For the major COG/KEGG categories (metabolism, cellular processes and signaling, and Information storage and processing), sub-classifications are shown in **C**, **D**, and **E**, respectively.

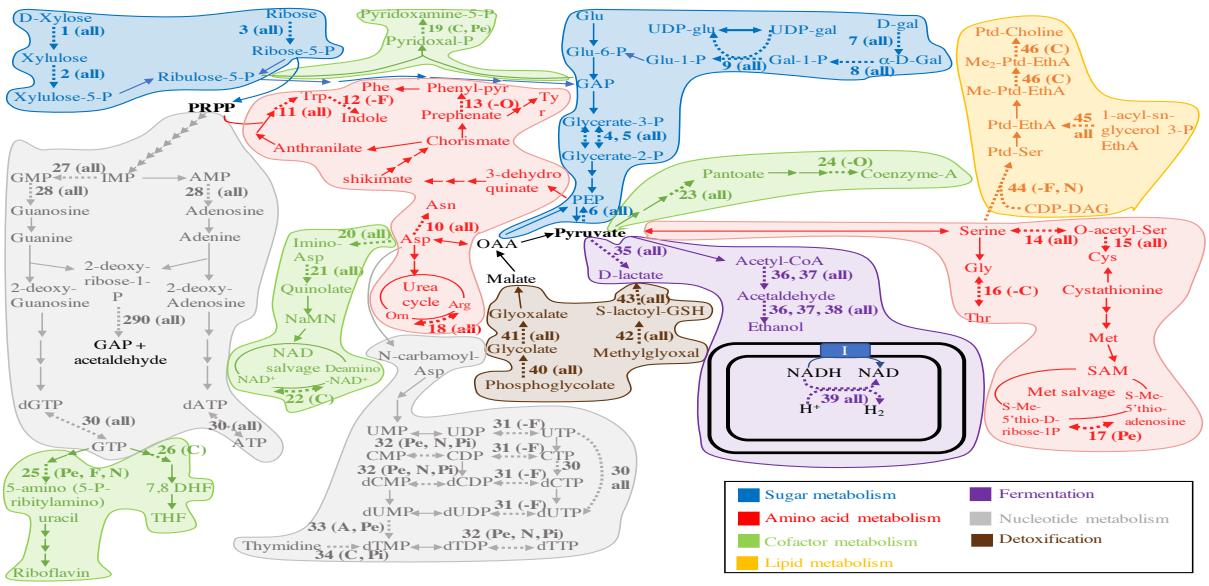
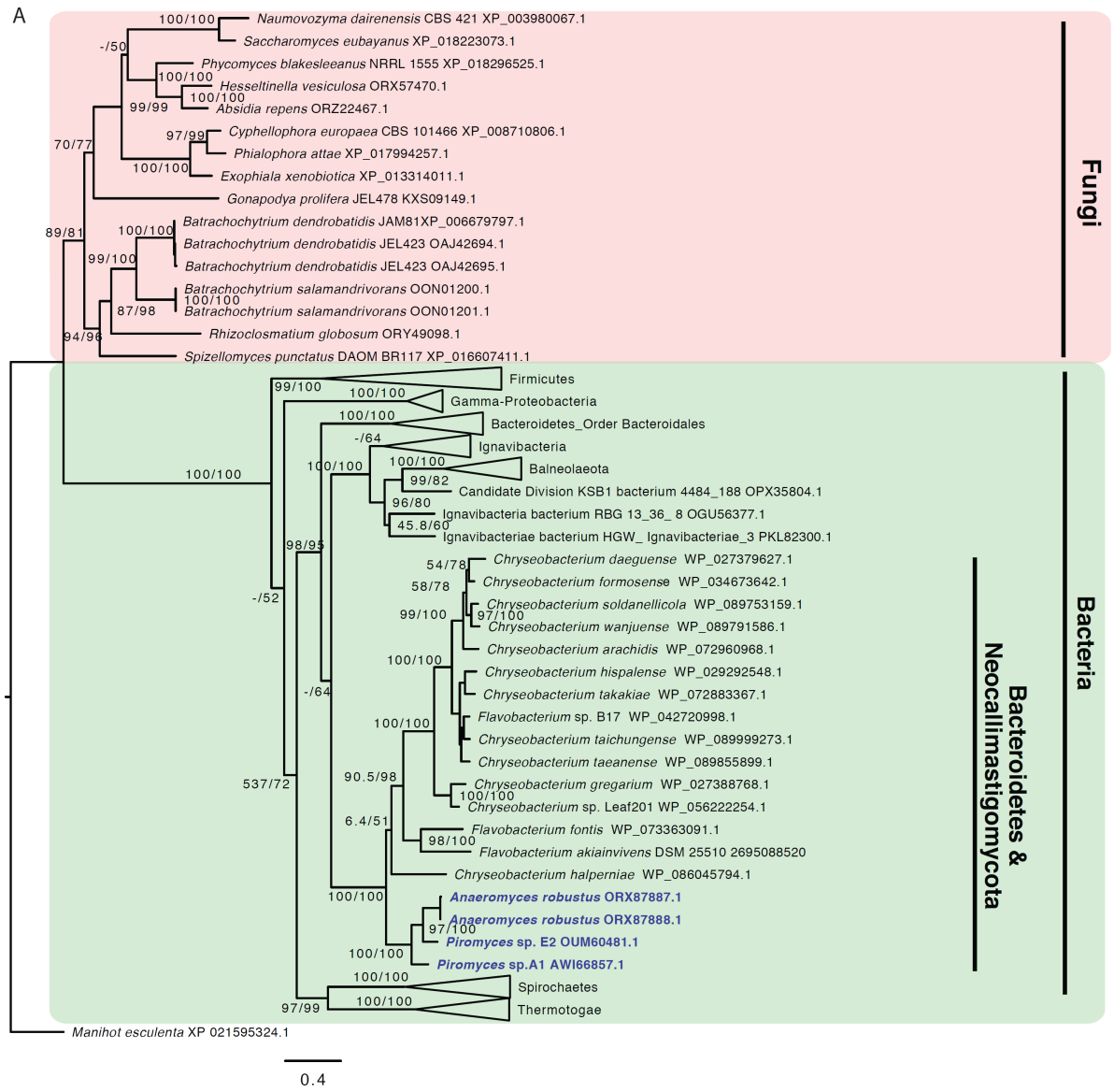


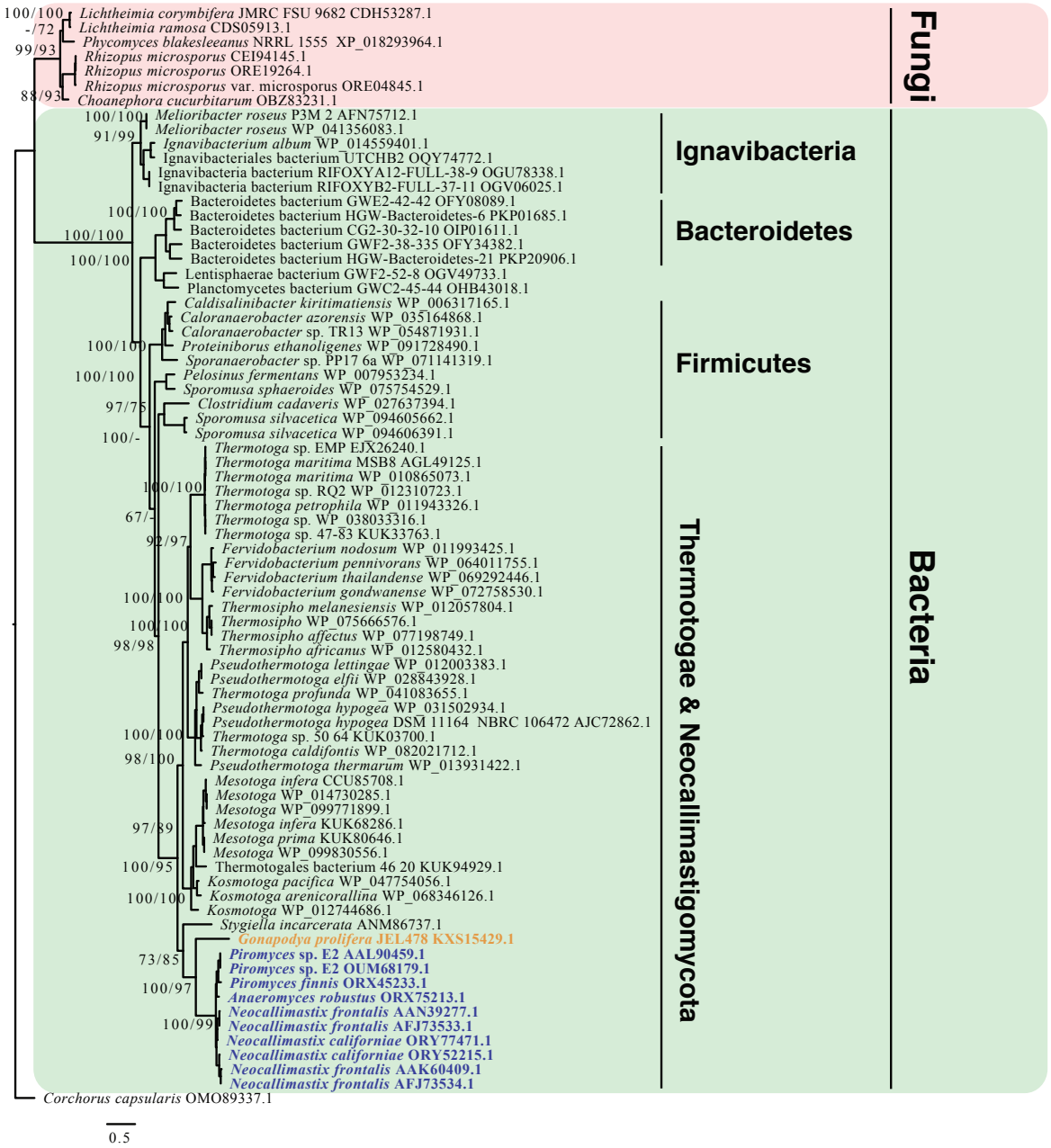

Figure 1.4: HGT impact on AGF central metabolic abilities. Pathways for sugar metabolism are highlighted in blue, pathways for amino acid metabolism are highlighted in red, pathways for cofactor metabolism are highlighted in green, pathways for nucleotide metabolism are highlighted in grey, pathways for lipid metabolism are highlighted in orange, fermentation pathways are highlighted in purple, while pathways for detoxification are highlighted in brown. The double black lines depict the hydrogenosomal outer and inner membrane. Arrows corresponding to enzymes encoded by horizontally transferred transcripts are shown with thicker dotted lines and are given numbers 1 through 46 as follows. Sugar metabolism (1-9): 1. Xylose isomerase, 2. Xylulokinase, 3. Ribokinase, 4. 2,3-bisphosphoglycerate-independent phosphoglycerate mutase, 5. 2,3-bisphosphoglycerate-dependent phosphoglycerate mutase, 6. Phosphoenolpyruvate synthase, 7. Aldose-1-epimerase, 8. Galactokinase, 9. Galactose-1-phosphate uridyltransferase. Amino acid metabolism (10-18): 10. Aspartate-ammonia ligase, 11. Tryptophan synthase (TrpB), 12. Tryptophanase, 13. Monofunctional prephenate dehydratase, 14. Serine-O-acetyltransferase, 15. Cysteine synthase, 16. Low-specificity threonine aldolase, 17. 5'-methylthioadenosine nucleosidase/5'-methylthioadenosine phosphorylase (MTA phosphorylase), 18. Arginase. Cofactor metabolism (19-26): 19. Pyridoxamine 5'-phosphate oxidase, 20. L-aspartate oxidase (NadB), 21. Quinolate synthase (NadA), 22. NH(3)-dependent NAD(+) synthetase (NadE), 23. 2-dehydropantoate 2-reductase, 24. dephosphoCoA kinase, 25. Dihydrofolate reductase (DHFR) family, 26. Dihydropteroate synthase. Nucleotide metabolism (27-34): 27. GMP reductase, 28. Trifunctional nucleotide phosphoesterase, 29. deoxyribose-phosphate aldolase (DeoC), 30. Oxygen-sensitive ribonucleoside-triphosphate reductase class III (NrdD), 31. nucleoside/nucleotide kinase family protein, 32. Cytidylate kinase-like family, 33. thymidylate synthase, 34. thymidine kinase. Pyruvate metabolism (fermentation pathways) (35-39): 35. D-lactate dehydrogenase, 36. bifunctional aldehyde/alcohol dehydrogenase family of Fe-alcohol dehydrogenase, 37. Butanol dehydrogenase family of Fe-alcohol dehydrogenase, 38. Zn-type alcohol dehydrogenase, 39. Fe-only hydrogenase. Detoxification reactions (40-43): 40. Phosphoglycolate phosphatase, 41. Glyoxal reductase, 42. Glyoxalase I, 43. Glyoxalase II. Lipid metabolism (44-46): 44. CDP-diacylglycerol–serine O-phosphatidyltransferase, 45. lysophospholipid acyltransferase LPEAT, 46. methylene-fatty-acyl-phospholipid synthase. Following the numbers, between parentheses, the distribution of the specific event across AGF genera is shown where (all) indicates the event was detected in all 7 genera, while a minus sign followed by a genus indicates that the event was detected in all but that/those genus/genera. Genera are represented by letters as follows: A, *Anaeromyces*; C, *Caecomyces*; F, *Feramyces*, N, *Neocallimastix*, O, *Orpinomyces*; Pe, *Pecoramyces*; Pi, *Piromyces*. Abbreviations: CDP-DAG, CDP-diacylglycerol; 7,8 DHF, 7,8 dihydrofolate; EthA, ethanolamine; Gal, galactose; GAP, glyceraldehyde-3-P; Glu, glucose; GSH, glutathione; I, complex I NADH dehydrogenase; NaMN, Nicotinate D-ribonucleotide; Orn, ornithine; PEP, phosphoenol pyruvate; Phenyl-pyr, phenylpyruvate; PRPP, phosphoribosyl-pyrophosphate; Ptd, phosphatidyl; SAM; S-adenosylmethionine; THF, tetrahydrofolate.

Figure 1.5: Phylogenetic trees

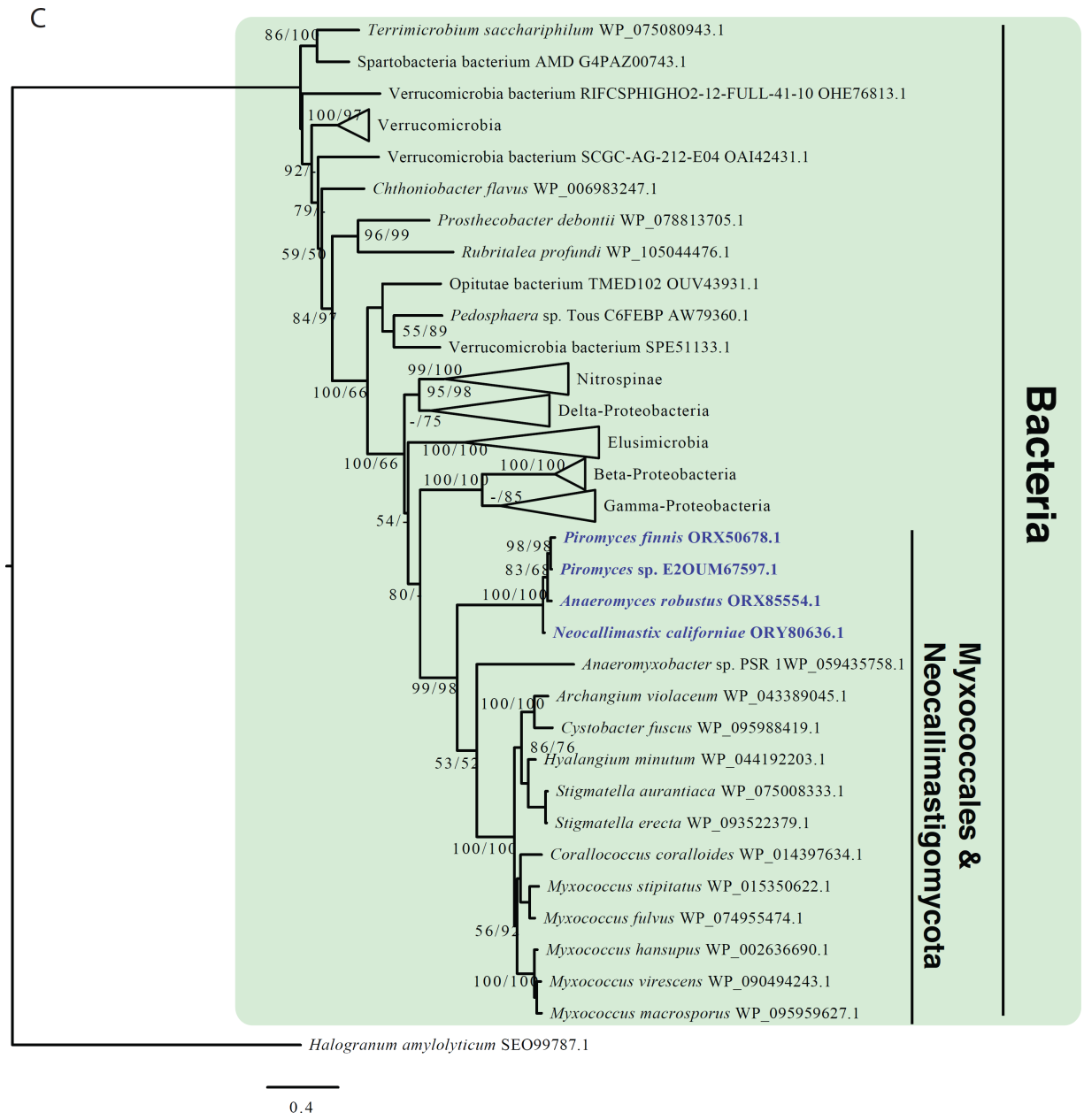



(a) Maximum likelihood tree showing the phylogenetic affiliation of AGF galactokinase. AGF genes highlighted in light blue clustered within the Flavobacteriales order of the Bacteroidetes phylum and were clearly nested within the bacterial domain (highlighted in green) attesting to their non-fungal origin. Fungal galactokinase representatives are highlighted in pink.

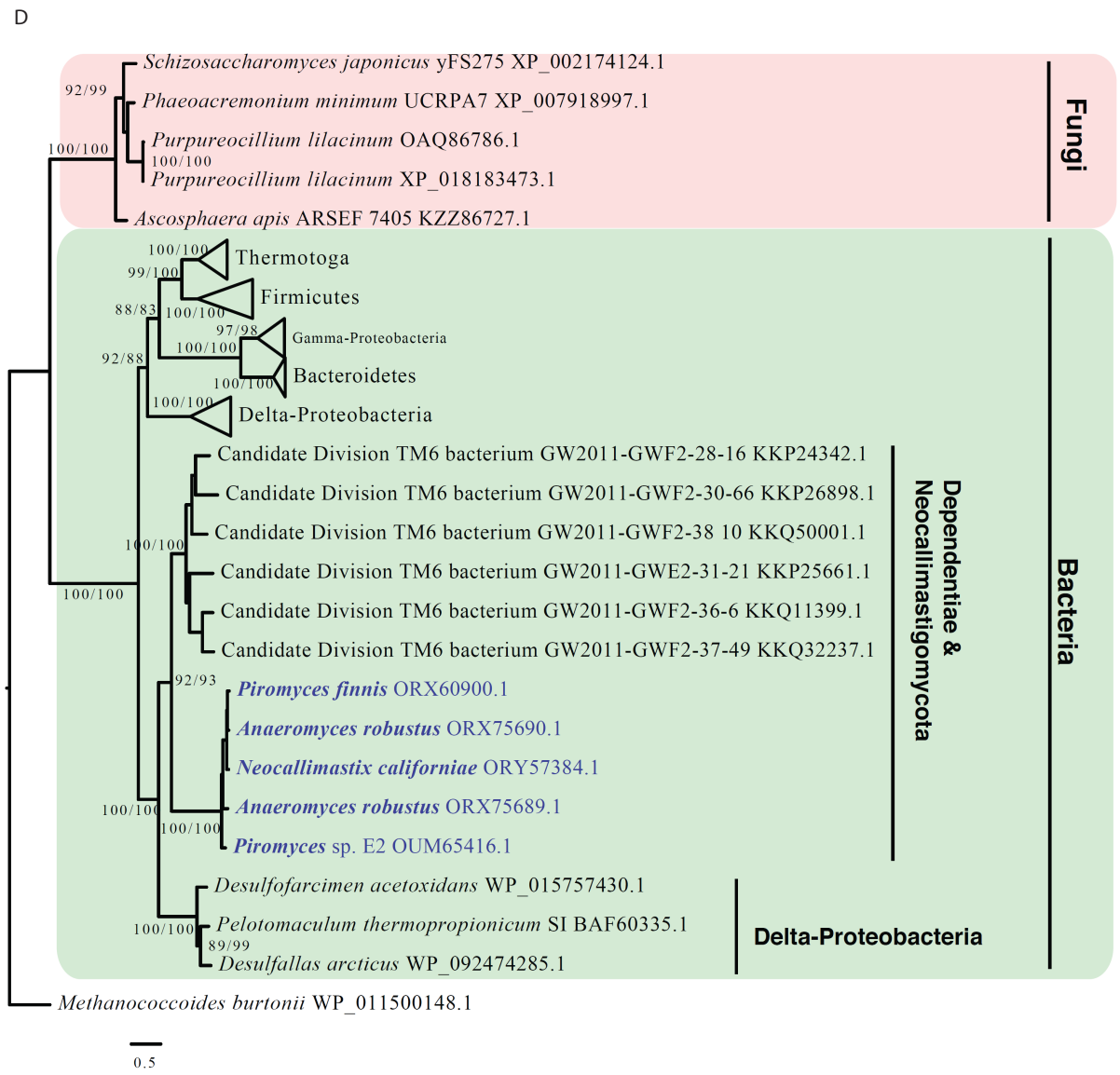
B




(b) Maximum likelihood tree showing the phylogenetic affiliation of AGF Fe-only hydrogenase. AGF genes highlighted in light blue clustered within the Thermotogae phylum and were clearly nested within the bacterial domain (highlighted in green) attesting to their non-fungal origin. Stygiella incarcerata (anaerobic Jakobidae) clustered with the Thermotogae as well, as has recently been suggested [229]. Fe-only hydrogenases from Gonopodya prolifera (Chytridiomycota) (shown in orange text) clustered with the AGF genes. This is an example of one of the rare occasions (n=24) where a non-AGF basal fungal representative showed an HGT pattern with the same donor affiliation as the Neocallimastigomycota. Other basal fungal Fe-only hydrogenase representatives are highlighted in pink and clustered outside the bacterial domain.




(c) Maximum likelihood tree showing the phylogenetic affiliation of AGF L-aspartate oxidase (NadB). AGF genes highlighted in light blue clustered within the Delta-Proteobacteria class and were clearly nested within the bacterial domain (highlighted in green) attesting to their non-fungal origin. As de-novo NAD synthesis in fungi usually follow the five-enzyme pathway starting from tryptophan, as opposed to the two-enzyme pathway from aspartate, no NadB were found in non-AGF fungi and hence no fungal cluster is shown in the tree.




(d) Maximum likelihood tree showing the phylogenetic affiliation of AGF oxygen-sensitive ribonucleotide reductase (NrdD). AGF genes highlighted in light blue clustered with representatives from Candidate phylum Dependentiae and were clearly nested within the bacterial domain (highlighted in green) attesting to their non-fungal origin. Fungal NrdD representatives are highlighted in pink. GenBank accession numbers are shown in parentheses. Alignment was done using the standalone MAFFT aligner [201] and trees were constructed using IQ-tree [281].

| Family | Genus | | | | | | |
|---|---|---|---|---|---|---|---|
| | *Anaeromyces* | *Caecomyces* | *Pecoramyces* | *Piromyces* | *Neocallimastix* | *Feramyces* | *Orpinomyces* |
| GH1 | 8 | | | | | | 4 |
| GH2 | 4 | | 4 | 4 | 4 | 4 | 4 |
| GH3 | 7, 15 | 7, 15 | 7, 15 | 7, 15 | 7 | | 7 |
| GH5 | 12 | 4, 12 | 4, 12 | 4, 12 | 4, 12 | 4, 12 | 4, 12 |
| GH6 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| GH8 | 3, 5 | 3 | 3, 5 | 3, 5 | 3, 5 | 5 | 3, 5 |
| GH9 | | | | | | | |
| GH10 | 4, 12 | 4, 12 | 4, 12 | 4, 12 | 4, 12 | 4, 12 | 4, 12 |
| GH11 | 3, 4 | 3, 4 | 3, 4 | 3, 4 | 3, 4 | 3, 4 | 3, 4 |
| GH13 | 2, 4, 6 | 2, 4, 6 | 2, 4, 6 | 2, 6 | 2, 4, 6 | 2, 4, 6 | 2, 4, 6 |
| GH16 | 2, 4 | | 2 | 4 | 2, 4, 13 | 2, 4 | 2, 4 |
| GH18 | | 12 | 12 | 12 | 12 | 12 | |
| GH20 | | | | | | | |
| GH24 | 4, 9 | | | 4 | | | 9 |
| GH25 | 4 | 4 | 4 | 4 | 4 | 4 | |
| GH26 | 3 | 3 | 3 | 3 | 3 | | |
| GH28 | | | 2 | | 2, 13 | | 2 |
| GH30 | 4 | | | | 4 | 4 | |
| GH31 | | | 1 | | | | |
| GH32 | 5 | | 5 | | 2, 13 | 5 | 5 |
| GH36 | | | | | | | |
| GH37 | | | | | | | |
| GH39 | 2, 4 | 4 | 2, 4 | 4 | 2, 4 | | 4 |
| GH43 | 4 | 4, 12 | 4 | 4, 8, 12 | 4, 12 | 4, 11 | 4 |
| GH45 | | | | | | | |
| GH47 | 15 | | 15 | 15 | 15 | 15 | |
| GH48 | 12 | 12 | 12 | 12 | 12 | 12 | 12 |
| GH53 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| GH57 | | | | | | | |
| GH64 | 4 | 4 | 4 | | 4 | 4 | |
| GH67 | 2 | 2 | 2 | 2 | | | |
| GH76 | | | | | | 2 | |
| GH78 | | | 12 | | | | |
| GH88 | | | 12 | 12 | 12 | 12 | 12 |
| GH95 | 2 | | | | | | 2 |
| GH97 | | | | | 3 | 3 | |
| GH108 | | | | 14 | | | |
| GH114 | | | | | | | |
| GH115 | 4 | 4 | 4 | 4 | 4 | 4 | |
| CE1 | 3, 4 | 3, 4 | 3, 4 | 3, 4 | 3, 4 | 3, 4 | 3, 4 |
| CE2 | 6, 12 | 6, 12 | 6, 12 | 6, 12 | 6, 12 | 2, 6 | 6 |
| CE3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| CE4 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| CE6 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| CE7 | 4 | | | | | | |
| CE8 | 13 | 13 | 13 | 13 | 13 | 13 | 13 |
| CE12 | 4, 12 | | 4 | 4, 12 | 4 | 2, 4 | 4 |
| CE15 | 12 | 3, 12 | 12 | 12 | 3, 12 | 12 | 12 |
| CE16 | | | | | 13 | | |
| PL1 | | | | | | | |
| PL3 | | | | | | | |
| PL4 | 12 | 12 | 12 | 12 | 12 | 12 | 12 |
| PL9 | 4, 11 | 4 | 4, 11 | 7, 11 | 4, 11 | | 11, 14 |
| PL11 | | | | | 4 | | |

| Donor Key | | | |
|---|---|---|---|
| 1 | Actinobacteria | | Bacteria |
| 2 | Bacteroidetes | | |
| 3 | Fibrobacter | | |
| 4 | Clostridiales | Firmicutes | |
| 5 | Lactobacillales | | |
| 6 | Unclassified Firmicutes | | |
| 7 | Lentisphaerae | | |
| 8 | Beta-Proteobacteria | Proteobacteria | |
| 9 | Gamma-Proteobacteria | | |
| 10 | Delta-Proteobacteria | | |
| 11 | Spirochaetes | | |
| 12 | Bacteria (unnested) | | |
| 13 | Viridplantae | | Eukaryota |
| 14 | Neagelaria | | |
| 15 | Unclassified Eukaryotes | | |

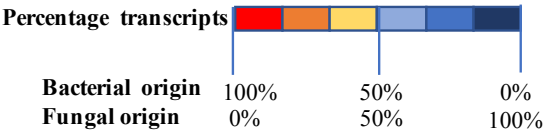



Figure 1.6: HGT in the AGF CAZyome shown across the seven genera studied. Glycosyl Hydrolase (GH), Carboxyl Esterase (CE), and Polysaccharide Lyase (PL) families are shown to the left. The color of the cells depicts the prevalence of HGT within each family. Red indicates that 100% of the CAZyme transcripts were horizontally transferred. Shades of red-orange indicate that HGT contributed to >50% of the transcripts belonging to that CAZy family. Dark blue indicates that 100% of the CAZyme transcripts were of fungal origin. Shades of blue indicate that HGT contributed to <50% of the transcripts belonging to that CAZy family. The numbers in each cell indicate the affiliation of the HGT donor as shown in the key to the right.

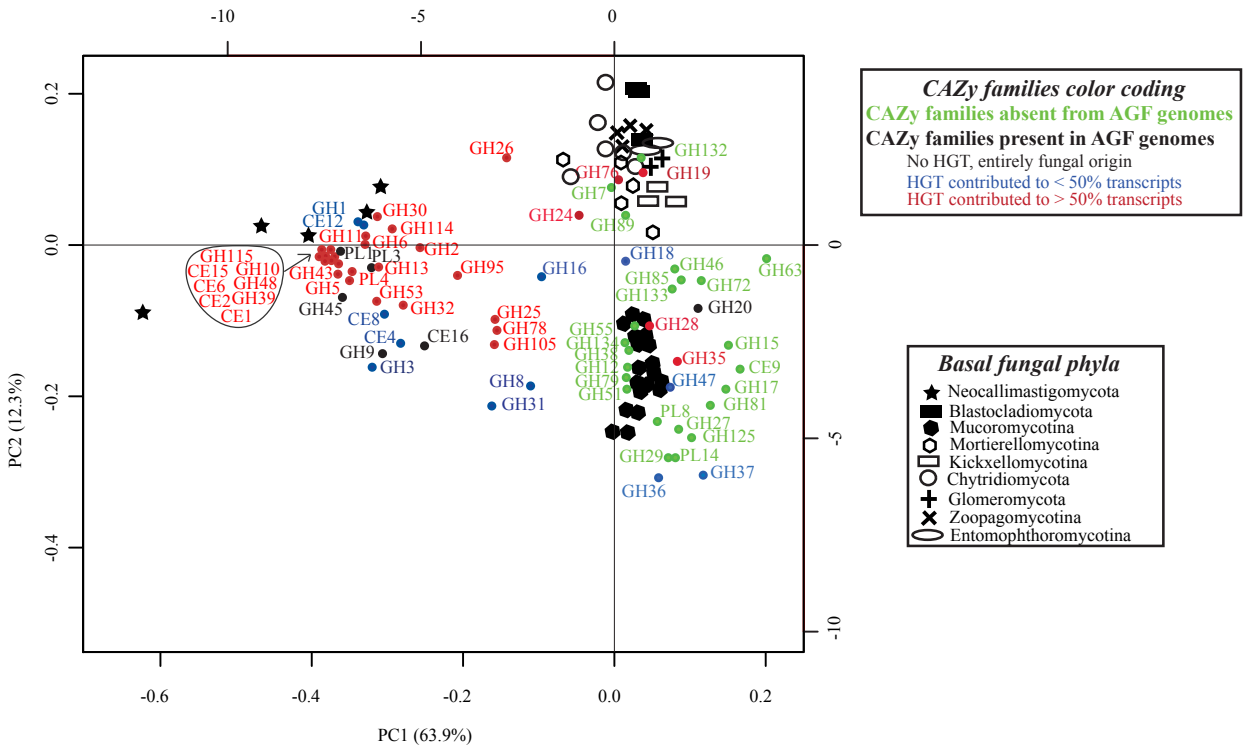



Figure 1.7: Principal-component analysis biplot of the distribution of CAZy families in AGF genomes (star), compared to representatives of other basal fungi belonging to the Mucoromycotina (dark hexagon), Chytridiomycota (circle), Blastocladiomycota (dark square), Entomophthoromycotina (oval), Mortierellomycotina (hexagon), Glomeromycota (plus), Kickxellomycotina (square), and Zoopagomycotina (X). CAZy families are shown as colored dots. The color code used was as follows: green, CAZy families that are absent from AGF genomes; black, CAZy families present in AGF genomes and with an entirely fungal origin; blue, CAZy families present in AGF genomes and for which HGT contributed to <50% of the transcripts in the examined transcriptomes; red, CAZy families present in AGF genomes and for which HGT contributed to >50% of the transcripts in the examined transcriptomes. The majority of CAZyme families defining the AGF CAZyome were predominantly of non-fungal origin (red and blue dots).

# CHAPTER II

# THE EXTRARADICAL PROTEINS OF *RHIZOPHAGUS IRREGULARIS*: A SHOTGUN PROTEOMICS APPROACH

## 2.1 Abstract

Arbuscular Mycorrhizal fungi (AMF, Glomeromycota) form obligate symbiotic associations with the roots of most terrestrial plants. Our understanding of the molecular mechanisms enabling AMF propagation and AMF-host interaction is currently incomplete. Analysis of AMF proteomes could yield important insights and generate hypotheses on the nature and mechanism of AMF-plant symbiosis. Here, we examined the extraradical mycelium proteomic profile of the arbuscular mycorrhizal fungus *Rhizophagus irregularis* grown on Ri T-DNA transformed Chicory roots in a root organ culture setting. Our analysis detected 529 different peptides that mapped to 474 translated proteins in the *Rhizophagus irregularis* genome. *R. irregularis* proteome was characterized by a high proportion of proteins (9.9 % of total, 21.4% of proteins with functional prediction) mediating a wide range of signal transduction processes, e.g. Rho1 and Bmh2, Ca-signaling (calmodulin, and Ca channel protein), mTOR signaling (MAP3K7, and MAPKAP1), and phosphatidate signaling (phospholipase D1/2) proteins, as well as members of the Ras signaling pathway. In addition, the proteome contained an unusually large proportion (53.6%) of hypothetical proteins, the majority of which (85.8%) were Glomeromycota-specific. Forty-eight proteins were predicted to be surface/membrane associated, including multiple hypothetical proteins of yet-unrecognized functions. However, no evidence for the overproduction of specific proteins, previously implicated in promoting soil health and aggregation was obtained. Finally, the comparison of *R. irregularis* proteome to previously published AMF proteomes identified a core set of pathways and processes involved in AMF growth. We conclude that *R. irregularis* growth on chicory roots requires the activation of a wide range of signal transduction pathways, the secretion of multiple novel hitherto unrecognized Glomeromycota-specific proteins, and the expression of a wide array of surface-membrane associated proteins for cross kingdom cell-to-cell communications.

## 2.2 Introduction

Arbuscular Mycorrhizal fungi (AMF, Glomeromycotina) are soil-dwelling fungi that form an obligate symbiotic association with the roots of ˜80% of terrestrial plant species [45, 358]. The AMF-plant root association represents one of the most significant symbiotic relationships in nature. In broad terms, the relationship is usually mutualistic and involves bidirectional transfers of fixed organic carbon from the plant and soil-derived nutrients from the fungi

[189]. AMF colonize the root cortex, where they penetrate and form arbuscules, coiled hyphae, and intraradical mycelium [98] inside the root cells. Outside the plant, extraradical mycelium extends meters away and plays important roles in nutrient (phosphorous, nitrogen, and zinc in particular) uptake and translocation to the intraradical structures [186]. Extraradical mycelia also play a role in foraging for new carbon sources and hence may associate with plants from the same or different species, forming an underground highway [358]. Anastomosis, the fusion of hyphae from the same or sometimes different fungal isolates belonging to the same species, is also frequently encountered [185]. This facilitates and maximizes nutrient uptake [185, 358].

Interest in various aspects of AMF biology has been fueled not only by their distinct phylogenetic position, unique evolutionary history, and obligate-biotrophic lifestyle, but also by their putative economic benefits, e.g. enhanced phosphorous supply to the plant [359, 360], increased resistance to drought [30, 38], enhanced disease resistance [166, 313, 361], improvement of soil structure [60, 255, 256] and fertility [1], and alleviation of metal toxicity [209, 339]. To-date, significant progress has been made in the area of AMF taxonomy using a combination of molecular/phylogenetic data and morphological characteristics [217, 320]. Similarly, the ecological distribution of AMF is currently fairly well characterized, with a plethora of studies characterizing the AMF diversity and community structure in specific ecosystems as well as on a global scale, and documenting AMF diversity in relation to the plant community, as well as in response to various soil management practices and anthropogenic perturbation [17, 18, 91, 109, 138, 154, 161, 168, 177, 213, 228, 230, 238, 260, 306, 334, 388, 406, 416].

Genomic and transcriptomic studies on AMF have lagged behind those of other fungi due to the obligate biotrophic nature of AMF, necessitating lengthy and laborious culturing approaches [49, 85, 127] for biomass acquisition, as well as the coenocytic nature (presence of multiple nuclei in a common cytoplasm) complicating the genome assembly process [67, 66, 206, 237, 243, 264, 373, 384]. Ttranscriptomic studies have been conducted on spores, extraradical mycelium, and symbiotic roots of *Rhizophagus intraradices* [383], and provided valuable insights on their symbiosis-associated traits and suggested the presence and possible operation of a cryptic sexual cycle. Similarly, transcriptomic sequencing of *Gigaspora margarita* in presence and absence of its endosymbiotic bacteria was instrumental in highlighting the role of the endobacteria in the pre-symbiotic phase of the fungus with its plant host [343].

With all the benefits of transcriptomics, translational and post-translational modifications often occur on secreted proteins, and hence the correlation between mRNA transcripts levels and expressed proteins concentrations is not always straightforward [120]. As such, proteomics factor as a promising approach in AMF research, since it provides a realistic picture of gene function [387]. Recent advances in shotgun proteomics and data analysis allowed for better detection and optimization of low abundance protein coverage, and overcame the underrepresentation of proteins with very low or very high pI and/or molecular weight that is known to occur with alternative methods, e.g. 2D-PAGE [156, 319]. As well, the recent availability of genomic sequences of AMF fungi (*Rhizophagus* and *Gigaspora* species [67, 66, 206, 237, 243, 264, 373, 384]) provides a better opportunity for positive identification of peptides based on a library. Here, we sought to exploit these improvements to investigate the extraradical proteomic profile of *Rhizophagus irregularis* grown on Ri T-

DNA transformed Chicory roots in a root-organ culture setting. In addition to providing a more detailed proteomic dataset than prior studies, we specifically sought to investigate: 1. Whether unique AMF fungal proteins are necessary for initiating biotrophic interaction between AMF and plant hosts, the identity of these proteins, and their putative conservation across AMF taxa, and 2. The putative role of AMF in maintaining and promoting organic carbon deposition in soil [155, 159, 395, 400] by producing copious amounts of extraradical cell wall proteins that forms the basis for additional particle aggregation and improvement of soil organic carbon content as previously proposed [326]. We provide a more comprehensive *Rhizophagus irregularis* extraradical mycelium (ERM) proteome, compare it to previous AMF shotgun proteomics studies, and suggest the presence of an ERM core proteome essential for interaction with the plant host.

## 2.3 Materials and Methods

### 2.3.1 Organism and growth conditions

*Rhizophagus irregularis* (Błaszkowski, Wubet, Renker & Buscot, C. Walker & A. Schüßler) spores were purchased from the Belgian Co-ordinated Collections of Microorganisms (BCCM) and used to start root-organ cultures on Ri T-DNA transformed chicory roots (*Chicorium intybus;* also purchased from BCCM) as described before [367]. Briefly, original root-organ cultures were started by inoculating transformed roots on modified Strullu Romand medium (MSR) solidified with 0.3% phytagel [95] with fungal spores at the root apex as well as at the intersection of the main root with new roots. Plates were incubated at 27$^{o}$C in the dark for 12-18 weeks and were routinely examined under a stereomicroscope for fungal colonization and production of spores and extraradical mycelium. Following this initial incubation, gel plugs containing abundant hyphae and spores were excised from the plates and used for subculturing on dual-compartment plates, with one compartment (root compartment) filled with MSR medium with sucrose, while the other compartment (hyphal compartment) lacking the carbon source. Plates were again incubated at 27$^{o}$C in the dark for a 12-week period, during which they were examined for mycorrhizal colonization. Roots were trimmed back to prevent their growth in the hyphal compartment. At the end of the incubation period, areas of abundant extraradical mycelium were excised from the plate and the medium was removed by incubation in 0.1M sodium citrate buffer (pH 6) for 1 hour at room temperature. Extraradical material were then collected by centrifugation, washed in sterile deionized water, ground in liquid nitrogen to a fine powder, and stored at -20$^{o}$C until further processed.

### 2.3.2 Protein extraction and SDS PAGE

Cold acetone wash was performed several times on the ground hyphal material followed by an overnight incubation with cold acetone at -20$^{o}$C. The pellet was then resuspended in SDS-PAGE Laemmli sample loading buffer and boiled for 3 minutes, after which the supernatant was loaded on a linear 12%, pH 8.8, SDS-PAGE gel and stained with Coomassie Blue. Each plate extracted yielded about 50 $\mu$g total protein, half of which were loaded on a single SDS-PAGE gel. Three separate growth and protein extraction experiments were conducted.

### 2.3.3 In-gel trypsin digestion, and LC-MS/MS

PAGE gels were processed by in-gel digestion [349] of the whole proteome displayed within the gel lanes [31, 172] following the detailed methodologies previously described [389]. Briefly, gel segments were excised, destained, reduced and alkylated, and infiltrated with trypsin/LysC solution (Promega). After overnight digestion, peptides were extracted from the gel digestions, extracts from all of the gel segments were pooled into a single fraction, and the peptides were purified further by solid phase extraction using monolithic C18 pipet tips (Agilent). Solution digests were performed using the FASP technique as described by [402]. Briefly, samples prepared in buffered SDS were subjected to a buffer exchange into 8M Tris-buffered urea containing TCEP and IAA (pH 8.5), then into 100 mM TrisHCl (pH 8.5) containing 5 $\mu$g/ml trypsin/LysC mix. Samples were digested overnight, after which an additional aliquot of trypsin/LysC was applied, and reactions were incubated 6 hr further. After digestion, peptides were collected by centrifugation, and purified by solid phase extraction using monolithic C18 pipet tips (Agilent). Purified peptides were dissolved in 0.1% formic acid, and injected onto a 75 $\mu$m x 20 mm C18 trap column (Fisher DX164535) connected to a 75 $\mu$m x 500 mm C18 analytical column (Fisher 164942) in a vented column configuration. An Easy-nLC 1200 UPLC system (Thermo) was used to elute the peptides from the column at 250 nL/min via a 4-37% gradient of mobile phase B (0.1:80:20 v/v/v formic acid/acetonitrile/water). During elution, peptides were ionized within a Nanospray Flex ion source (Thermo) via 1900 V applied to the stainless-steel emitter at the column terminus. Peptide ions were analyzed within a Fusion quadrupole hybrid mass spectrometer (Thermo), via a top-speed data-dependent MS/MS method wherein precursor ions were measured in the FT Orbitrap sector, subsequently isolated by quadrupole filtering, and fragmented by HCD activation, after which fragment ions were measured within the ion trap sector. Details regarding the specific instrument settings employed are in Table S1.

### 2.3.4 Proteins identification

Resultant mass spectrometry data were used to search a database consisting of *Rhizophagus irregularis* protein sequences downloaded from GenBank (GenBank taxonomy ID: 588596; retrieved in February 2018 and comprising 232,995 proteins), and also containing 176 *Chicorium intybus* sequences downloaded from Uniprot (February 2018). The searches also encompassed the sequences of decoy proteins (reversed sequences from the primary database) and common laboratory protein contaminants. Searches were performed using the Andromeda application embedded in MaxQuant v1.5.3.8 [84]. In addition to the default MaxQuant settings, searches utilized cleavage with trypsin/P tolerating 3 missed cleavages, and the following variable peptide modifications: oxidation of Met, carbamidomethylation of Cys, propionamide modification of Cys, Gln cyclization to pyroGlu, and carbamylation of Arg, Lys, and peptide N-termini. The MaxQuant second peptides search feature was also enabled. Data files from 8 separate LC-MS/MS analyses were concatenated into one MaxQuant experiment to generate the list of *Rhizophagus irregularis* proteins identified in the study (Table S2). Peptide identifications were accepted at a 1% false discovery rate, corresponding to a posterior error probability threshold of 0.016 or better. FDR filtering via the target-decoy approach is an efficient means to reduce misidentifications by modified peptides. (Bog-

danow et al., 2016) Protein identifications were accepted at a 1% false discovery rate, on the basis of one or more identified peptides [144]. The dependence on a one-peptide rather than a two-peptide rule for protein identification is based on the findings of [144]. In this study, a systematic evaluation has revealed that the dogmatic two peptide rule often reduces the number of protein identifications in the target database more significantly than in the decoy database; resulting in an increase in false discovery rates. Thus, the alternative "one-peptide" rule is well justified in our current work, especially given the biochemical challenges of the *R. irregularis* system. Raw data are presented in Supplementary Tables S2-S5.

### 2.3.5 Sequence analysis

Identified proteins were functionally classified using both KEGG database using BlastKOALA [197] as well as the Conserved Domain Database [246] using the COG database. Output from BlastKOALA was used as an input to the Search&Color Pathway module in KEGG mapper [199] to retrieve biochemical pathways. The Compute p$I$/Mw tool from ExPASy [136] was used to calculate theoretical molecular weight (Mw) and isoelectric point (p$I$). Subcellular location of proteins was predicted using PSORT II [277]. Transmembrane (TM) domains were predicted using the TMpred server (embnet.vital-it.ch/software/TMPRED_form.html).

### 2.3.6 Specificity of hypothetical proteins to the phylum Glomeromycota, the genus *Rhizophagus*, and the species *R. irregularis*

To identify which hypothetical proteins are Glomeromycota specific, genus *Rhizophagus* specific, or *R. irregularis* specific, we queried all identified proteins with an unknown function against the nr database using web Blastp [54](Camacho et al., 2009)[77] in a three-tier approach. First, to identify Glomeromycota specific proteins, all hypothetical proteins were queried against the Fungi (GenBank taxonomy ID: 4751) excluding the Glomeromycota (GenBank taxonomy ID: 214504). A hit was defined based on e-value ($e^{-10}$ cutoff) and alignment length (at least 100 aa). Proteins were deemed Glomeromycota-specific if they have no fungal hits outside the Glomeromycota. Second, to identify *Rhizophagus*-specific proteins, all Glomeromycota-specific proteins from the first tier were queried against the Glomeromycota (GenBank taxonomy ID: 214504) excluding the genus *Rhizophagus* (GenBank taxonomy ID: 1129544). Proteins were deemed *Rhizophagus*-specific if they have no hits outside the genus. Third, to identify *Rhizophagus irregularis*-specific proteins, all *Rhizophagus*-specific proteins from the second tier were queried against the genus *Rhizophagus* (GenBank taxonomy ID: 1129544) excluding the species *Rhizophagus irregularis* (GenBank taxonomy ID: 588596), and proteins with no hits were considered species specific.

### 2.3.7 Comparative analysis to previous studies of the ERM proteome of AMF

Aiming to draw a more comprehensive picture of the core proteins and pathways activated in the ERM proteome of AMF species, we compared the proteins predicted in this study to those obtained in previous studies that employed LC/MS/MS to analyze the ERM of *Rhizophagus* [319], and *Gigaspora* [387]. Using standalone BLAST+ suite, an all versus all Blastp analysis was conducted where predicted proteins from [319], [387], and this study

were compared. A hit was defined based on e-value ($e^{-10}$ cutoff) and alignment length (at least 100 aa).

### 2.3.8 Data Accession

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [1] partner repository with the dataset identifier PXD016091.

## 2.4 Results and Discussion

### 2.4.1 Protein identification

LC-MS/MS analysis of *R. irregularis* proteome detected 529 different peptides that were confidently mapped to 474 *Rhizophagus irregularis* protein sequences using the Andromeda search engine against a *Rhizophagus irregularis* GenBank database (GenBank taxonomy ID: 588596; retrieved in February 2018 and comprising 232,995 proteins). MaxQuant peptide spectrum match (PSM) scores for the detected peptides ranged from 40.1-186, and the posterior error probability ranged from $3.4x10^{-12}$ to 0.0165. Identified proteins spanned a range of hypothetical average pI (3.7-12.3) (Figure 1a), hypothetical molecular weights (6.4-567.8 KDa) (Figure 1b), and subcellular predicted locations (Figure 1c).

### 2.4.2 Functional significance of the identified proteins

A predicted function was identified for 220 proteins within the 474-protein dataset. The majority of the identified proteins with function prediction in this study were assigned to cellular processes and signaling (41.4%) and information storage and processing (36.4%); with only 9.1% of proteins involved in metabolic processes (Table 1, Figure 2).

#### Functions assigned to cellular processes and signaling

Ninety-two proteins (19.4% of total proteins, 41.8% of proteins with known functions) were assigned to the COG group "Cellular processes and signaling", with the majority belonging to the COG categories T (signal transduction, 47 proteins), and O (Post-translational modification, protein turnover, chaperone functions, 22 proteins).

Forty-seven proteins (9.9% of total, 21.4% of proteins with known function) mediating signal transduction (COG category "T") was identified. Identified signal transduction proteins include Rho1 and Bmh2 proteins [319], members of Ca-signaling (calmodulin, and Ca channel protein), mTOR signaling (MAP3K7, and MAPKAP1), and phosphatidate signaling (phospholipase D1/2) proteins, as well as members of the Ras signaling pathway (Table 1). The remaining proteins possessed domains implicated in signal transduction (e.g. protein tyrosine kinase domain, Sel1 repeat, and Tetratricopeptide repeat) were also identified (Table 1). The relative proportion of signal transduction proteins within *R. irregularis* proteome is similar to those identified within host-associated fungi (e.g. 18.2% in the pathogenic *Mucor circinelloides* [81]), and much higher than values reported in non-host associated fungi (for example, only 10.7% of the annotated genes in *Neurospora crassa* are predicted to have a signal transduction function [81]), suggesting the importance of signal transduction for

AMF biotrophic growth. Unfortunately, the obligate biotrophic nature of AMF precludes the use of a no-plant control to decipher which functions are important for cross-kingdom communication. However, the fact that signal transduction proteins were identified in ERM proteomes with smaller sizes [319, 387] hints at their importance in fungi-host communication.

Proteins affiliated with the COG category "O" included chaperones (heat shock protein Hsp70, and Hsp90), subunits of the chaperonin-containing T-complex, and the HSP70 co-chaperone sacsin, in addition to prefoldin subunits, important for proper folding of newly synthesized polypeptides. Proteins involved in polypeptide turnover, e.g. ubiquitin-conjugating and ubiquitin-activating enzymes, ubiquitin-protein ligase, and ubiquitin carboxyl terminal hydrolase, as well as several subunits of the proteasome were also identified.

Beside signal transduction (COG category "T") and protein folding and turnover (COG category "O"), other cellular processing functions identified include intracellular trafficking and vesicular transport believed to be essential for polar growth of filamentous fungi. These included: 1. Ras-related proteins and ADP-ribosylation factor, both involved in vesicular transport between the Golgi apparatus and the endoplasmic reticulum, 2. Subunits of importin for protein nuclear import, and 3. components of the PAN1 actin cytoskeleton regulatory complex involved in endosome internalization. Complementary to that, we also identified kinesin motor proteins known to be involved in anterograde transport of proteins and membrane components.

### Functions assigned to information storage and processing

Eighty proteins were assigned a function related to the COG group "Information storage and processing", with the majority belonging to the COG categories J (Translation; 26 proteins), K (Transcription; 19 proteins), L (Replication and DNA repair; 11 proeins), and X (Mobilome; 10 proteins). Functions identified include several transcription factors, proteins involved in both ribosome biogenesis (ribosomal biogenesis protein LAS1, nucleolar pre-ribosomal-associated protein 1, U3 small nucleolar RNA-associated protein 23), as well as the integral components of the translational machinery (translation initiation factors 2C, 3A, and 4A, translation elongation factors 1-alpha, and 2, peptide chain release factor subunit 1, several ribosomal proteins, tRNA synthetases, and tRNA ligase), DNA repair and recombination-related proteins (DNA helicases INO80 and MCM8, DNA polymerase gamma 1 and family B, DNA repair and recombination protein RAD54, initiator replication protein , and flap endonuclease-1), several chromatin remodeling proteins including the previously reported histones [319], and several RNA transport and processing proteins including several components of the spliceosome. Surprisingly, several transposases and reverse transcriptases were also identified.

### Functions assigned to metabolism

Twenty proteins assigned a function related to the COG group "metabolism" were identified. Proteins identified were enzymes belonging to a wide range of catabolic pathways (glycolysis, TCA cycle, $\beta$-oxidation, and oxidative phosphorylation). In addition, we identified proteins involved in the biosynthesis in the cofactors biotin and pantothenate, and an enzyme involved

in carnitine biosynthesis. Carnitine is known to facilitate the transfer of long-chain acyl-CoA molecules into the mitochondria for their use in $\beta$-oxidation [374].

**Hypothetical proteins**

A total of 254 *Rhizophagus* proteins (53.6% of total identified proteins) had no assigned function. This value is similar to the percentage of proteins with hypothetical functions in *Rhizophagus* reference proteome (47.5%; 70,195 proteins out of a total of 147,766 proteins in the Uniprot *Rhizophagus* reference proteome), but is much higher than the percentage of predicted hypothetical proteins in other fungal genomes (e.g. 21.9% of Aspergillus Uniprot reference proteome is comprised of proteins with hypothetical functions). The majority of these hypothetical proteins are predicted to be cytoplasmic (79.9%), with 12.2% predicted to be surface or membrane proteins, and only 7.9% predicted to be mitochondrial. Remarkably, the absolute majority (218 protein; 85.8%) of these hypothetical proteins were specific to the phylum Glomeromycota with no hits outside the phylum. Of these, 97 were specific to the genus *Rhizophagus*, and, of these, 54 were specific to the species *R. irregularis.*

### 2.4.3 Surface and membrane associated proteins in *R. irregularis* ERM proteome: Implications on production of glomalin-related surface proteins

It has previously been assumed that AMF produce copious amounts of cell wall proteins (collectively known as glomalin-related surface protein; GRSP) that sloughs off to the soil and form the basis for additional particle aggregation and improvement of soil organic carbon content [326]. It is worth mentioning that these studies used monoclonal antibodies raised against whole crushed spores of *Rhizophagus intraradices* [403], and hence there is no guarantee that they target a specific, well defined Glomeromycota membrane protein(s). In fact, the specificity of the monoclonal antibodies has been called into question [335] due to cross reactivity to various proteins, e.g. exogenously added bovine serum albumin, and carbon compounds from leaf litter.

Only 15 of the 48 surface/transmembrane proteins identified in this study had a predicted function. Of these, ten are predicted to have a cell membrane-associated function or are destined to other membranes in the cell (e.g. nucleus, vacuole) and hence could be excluded as possible candidates for GRSP due to their location (Table 1). Four are predicted to be membrane proteins involved in post-translational modification of proteins (glycosylation, phosphorylation, and O-acylation, as well as an aspartyl protease function). The final protein is predicted to have a chaperone function (HSP70). It is interesting to note that prior work have proposed an alternative function of GRSP as chaperonins (heat shock proteins) [128, 312]. The remaining 33 surface/transmembrane proteins had no predicted function (hypothetical proteins). Of these, 25 were specific to the phylum Glomeromycota.

GRSP are supposed to be produced in large quantities and accumulate to high levels (several folds of cytoplasmic proteins) on the cell surface. However, our peptide count data do not suggest that any of the 48 transmembrane/surface proteins are produced in exceptionally high levels. Accordingly, the current study does not support the notion that a large quantity of AMF surface proteins accumulates on the AMF hyphae. Further, the highly specialized predicted functions of most identified membrane proteins (e.g. post translational

modification) argues against their secretion in large quantities to the hyphal cell surface. It is possible that the production of these GRSP occurs only in soil, but not in axenic cultures. Nevertheless, the results presented support the opinion that GRSP are artifacts of prior inaccurate quantification procedures.

### 2.4.4 Comparative analysis to prior AMF ERM proteome studies

Previous studies employing LC/MS/MS to obtain and analyze the extraradical mycelium proteome of *Rhizophagus* [319], and *Gigaspora* [387] provided an excellent framework for understanding the intricate relationship between AMF and their plant host (in cases of *Rhizophagus* and *Gigaspora*), or AMF, their plant host and co-existing endobacteria (in case of *Gigaspora*). Ninety-two *Rhizophagus intraradices* ERM proteins were confidently assigned in [319]. Similarly, 156 *Gigaspora margarita* ERM proteins were confidently assigned in [387]. However, the absence of a robust database severely limited peptide identification in these two studies. The recent availability of *Rhizophagus irregularis* genome and transcriptome data allowed a significant improvement to the size of *Rhizophagus irregularis* ERM proteome identified in the current study.

Building on these previous studies, we aimed at drawing a more comprehensive picture of core proteins and pathways activated in ERM proteome of various AMF species. To this end, we performed an all versus all Blastp analysis to compare the predicted proteins from [319], [387], as well as the current study, and identified a set of functions that seem to be common at least to the two genera *Rhizophagus* and *Gigaspora.*

Metabolically, studying the ERM proteome of *Rhizophagus* and *Gigaspora* confirmed the functionality of the following pathways in all three proteomic studies analyzed: 1. Fatty acid beta-oxidation, TCA cycle, and oxidative phosphorylation for energy production, 2. Gluconeogenesis for production of hexose sugars, and the pentose phosphate pathway for production of NADPH and pentose sugars. 3. Chromatin remodeling, DNA replication and repair, protein synthesis (translation initiation and elongation factors, ribosomal proteins, and tRNA ligases) as evidence of active growth, and 4. Protein turnover within the proteasome via multiple proteases. Additionally, several families of chaperones and chaperonins were common members in all three ERM proteomes analyzed, including heat shock proteins (HSP60, 70, 80, and 90) and the chaperonin-containing T-complex. Further, cytoskeleton functions are also paramount in the extraradical mycelium as would be expected during polar growth. Oxidative stress response functions are also active to combat reactive oxygen species that accumulate during oxygen consumption. Finally, the adaptor protein 14-3-3 is a common member of *Rhizophagus* and *Gigaspora* ERM proteome and is probably regulating several signaling pathways for interaction with the plant host.

Quantitatively, a higher proportion of signal transduction proteins (21.4% of proteins with functional annotation, 9.9% of total proteins) were identified in this study as opposed to previous AMF proteome studies (only 10.9% in [319], and 3.8% in [387]). The proportion of surface/membrane proteins identified in this study (10.13% of total proteins) was similar to previous AMF proteome studies (10.9% in [319], and 8.97% in [387]). Given the difficulty of identification of hypothetical proteins in absence of reference genomes, it is difficult to ascertain whether the pattern of expression of a large number of hypothetical proteins was also occurring in *R. intraradices* [319] and *G. margarita* [387] proteomes.

### 2.4.5 Summary

In conclusion, our results suggest that *R. irregularis* growth on chicory roots requires the activation of a wide range of signal transduction pathways. Our data also do not lend support to prior hypothesis on the role of AMF in promoting soil health through secretion of GRSP. Finally, the identification of a wide range of hypothetical proteins with yet-unknown function clearly demonstrates that the molecular mechanisms underlying AMF growth is far from adequately understood.

## 2.5 Acknowledgements

This work was supported by the National Science Foundation Grant number 1649441.

This work is published and available in full online (DOI: 10.1016/j.funbio.2019.12.001).

## 2.6 Figures & Tables

Supplementary Figures and Tables can be viewed online at:
https://www.sciencedirect.com/science/article/pii/S1878614619301680?via%3Dihub

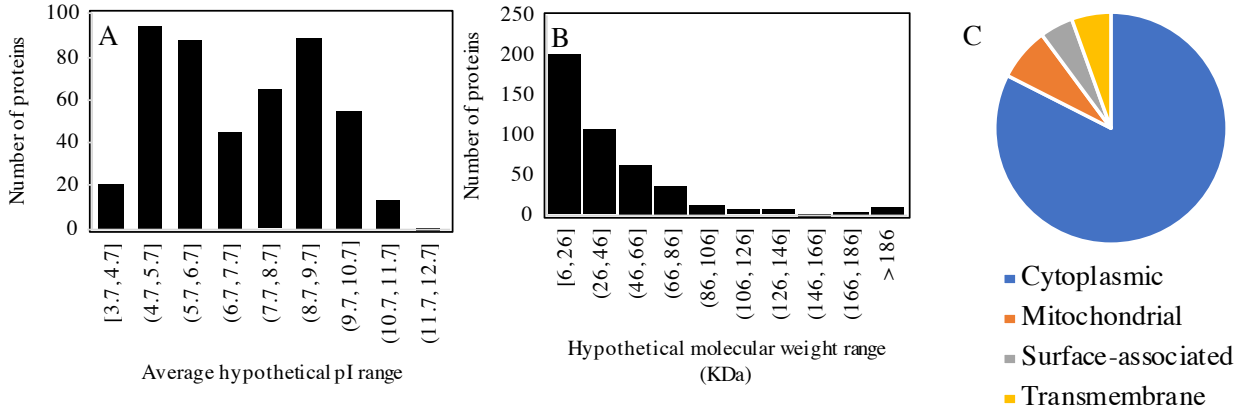



Figure 2.1: Physiological properties of the 474 proteins predicted in this study. Shown are histograms of the distribution of average hypothetical pI values **(A)**, and hypothetical molecular weight **(B)** predicted using the Compute pI/Mw tool from ExPASy [136], and a pie chart of the distribution of protein subcellular localization **(C)** predicted using PSORT II [277], and the TMpred server (https://embnet.vital-it.ch/software/TMPRED_form.html). Proteins predicted by PSORT II to be destined to the cell surface are depicted in **(C)** as "Surface-associated", while those predicted by TMpred to have transmembrane domain(s) are depicted in **(C)** as "Transmembrane"

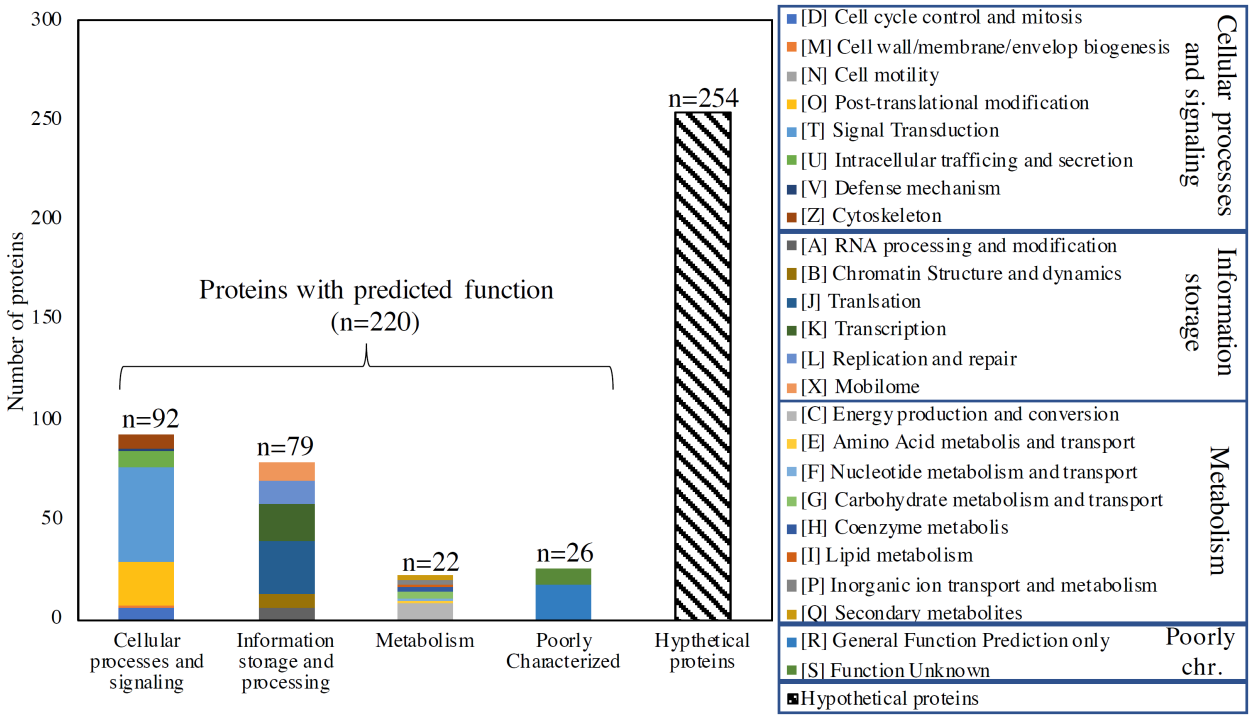



Figure 2.2: Functional annotation of the 474 proteins predicted in this study. The number of proteins is shown on the Y-axis. Proteins with functional COG categories are shown using colored bars, whereas hypothetical proteins with no functional annotation are shown separately on the right. The color legend is grouped by the COG categories belonging to the groups "Cellular processes and signaling", "Information storage and processing", "Metabolic processes", and "Poorly characterized". The total number of proteins belonging to each of these groups (n) is shown on top of the bars. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 2.1: List of the 220 proteins with predicted function that were identified in this study.

Table 1. List of the 220 proteins with predicted function that were identified in this study.

| GenBank Accession number | COG category | Predicted function |
|---|---|---|
| **RNA transport and processing** | | |
| EXX71251.1 | A | RAN; GTP-binding nuclear protein Ran |
| EXX59138.1 | A | GLE1; nucleoporin GLE1 |
| GBC34568.1 | A | DHX38; pre-mRNA-splicing factor ATP-dependent RNA helicase DHX38/PRP16 [EC:3.6.4.13] |
| POG74787.1 | A | DHX8; ATP-dependent RNA helicase DHX8/PRP22 [EC:3.6.4.13] |
| PKY21049.1 | A | SF3B1; splicing factor 3B subunit 1 |
| POG64456.1 | A | SF3B3; splicing factor 3B subunit 3 |
| **Chromatin structure and function** | | |
| PKY39409.1 | B | H2A; histone H2A |
| POG80393.1 | B | H4; histone H4 |
| PKY46020.1 | B | SMC1; structural maintenance of chromosome 1 |
| POG79749.1 | B | Myb-like DNA-binding domain |
| GBC30544.1 | B | Myb/SANT-like DNA-binding domain |
| EXX59487.1 | B | FACT complex subunit (SPT16/CDC68) |
| GBC35226.1 | B | HMG (high mobility group) box |
| **Metabolism- Energy production** | | |
| POG65659.1 | C | QCR2; ubiquinol-cytochrome c reductase core subunit 2 |
| EXX55638.1 | C | ATPeF1A; F-type H+-transporting ATPase subunit alpha |
| POG79600.1 | C | ATPeF1B; F-type H+-transporting ATPase subunit beta [EC:7.1.2.2] |
| POG74762.1 | C | ATPeV1A; V-type H+-transporting ATPase subunit A [EC:7.1.2.2] |
| POG78227.1 | C | CYC; cytochrome c |
| POG63694.1 | C | GRHPR; glyoxylate/hydroxypyruvate reductase [EC:1.1.1.79 1.1.1.81] |
| PKK70634.1 | C | FAD binding domain |
| EXX79393.1 | C | adh2; alcohol dehydrogenase [EC:1.1.1.-] |
| **Cell cycle control** | | |
| PKK77139.1 | D | Putative zinc-finger domain |
| POG76118.1 | D | Inhibitor of Apoptosis domain |
| EXX67272.1 | D | AIG1 family |
| PKY51503.1 | D | AIG1 family |
| POG83207.1 | D | TPX2; targeting protein for Xklp2 |
| PKK57576.1 | D | Antagonist of mitotic exit network protein 1 |
| **Amino acid metabolism** | | |
| EXX77353.1 | E | IVD; isovaleryl-CoA dehydrogenase [EC:1.3.8.4] |
| **Nucleotide transport** | | |
| EXX66135.1 | F | SLC25A4S; solute carrier family 25 (mitochondrial adenine nucleotide translocator), member 4/5/6/31 |
| **Carbohydrate metabolism** | | |
| PKC51795.1 | G | deoC; deoxyribose-phosphate aldolase [EC:4.1.2.4] |
| GBC34503.1 | G | ENO; enolase [EC:4.2.1.11] |
| GBC30498.1 | G | GAPDH; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] |
| EXX71892.1 | G | IDH1; isocitrate dehydrogenase [EC:1.1.1.42] |
| **Cofactor metabolism** | | |
| PKC12604.1 | H | bioB; biotin synthase [EC:2.8.1.6] |
| POG65272.1 | H | PPCS; phosphopantothenate---cysteine ligase (ATP) [EC:6.3.2.51] |
| **Lipid metabolism** | | |
| EXX75213.1 | I | E1.3.3.6; acyl-CoA oxidase [EC:1.3.3.6] |
| PKK80127.1 | I | PCCA; propionyl-CoA carboxylase alpha chain [EC:6.4.1.3] |
| **Translation** | | |
| **Ribosome biogenesis** | | |
| EXX53919.1 | J | DHX37; ATP-dependent RNA helicase DHX37/DHR1 [EC:3.6.4.13] |
| PKY38878.1 | J | LAS1; ribosomal biogenesis protein LAS1 |
| GBC41625.1 | J | URB1; nucleolar pre-ribosomal-associated protein 1 |
| POG67021.1 | J | UTP23; U3 small nucleolar RNA-associated protein 23 |
| **Protein synthesis** | | |
| PKK61018.1 | J | EEF1AKMT1; EEF1A lysine methyltransferase 1 [EC:2.1.1.-] |
| POG71413.1 | J | EIF3A; translation initiation factor 3 subunit A |
| PKK72370.1 | J | EEF1A; elongation factor 1-alpha |
| GBC21201.1 | J | EEF2; elongation factor 2 |
| PKC74556.1 | J | EIF4A; translation initiation factor 4A |
| EXX59520.1 | J | ELF2C; eukaryotic translation initiation factor 2C |
| PKC14039.1 | J | ELF2C; eukaryotic translation initiation factor 2C |
| PKY33432.1 | J | ELF2C; eukaryotic translation initiation factor 2C |
| POG69992.1 | J | ETF1; peptide chain release factor subunit 1 |
| PKC65668.1 | J | LARS; leucyl-tRNA synthetase [EC:6.1.1.4] |
| PKY39583.1 | J | Mitochondrial ribosomal protein mL59 |

| | | |
|---|---|---|
| EXX54999.1 | J | Ribosomal protein L10 |
| PKY18273.1 | J | Ribosomal protein L17 |
| POG71605.1 | J | RP-L27e; large subunit ribosomal protein L27e |
| PKY47892.1 | J | RP-L40e; large subunit ribosomal protein L40e |
| POG75786.1 | J | RP-S14e; small subunit ribosomal protein S14e |
| POG77340.1 | J | RP-S20e; small subunit ribosomal protein S20e |
| POG76532.1 | J | RP-S23e; small subunit ribosomal protein S23e |
| POG80381.1 | J | RP-S3e; small subunit ribosomal protein S3e |
| POG72388.1 | J | RP-S7e; small subunit ribosomal protein S7e |
| POG79986.1 | J | Threonyl and Alanyl tRNA synthetase second additional domain |
| POG78748.1 | J | TRL1; tRNA ligase [EC:6.5.1.3] |
| PKY50373.1 | JK | KH domain |
| **Transcription** | | |
| **Transcription factors** | | |
| PKY27414.1 | K | E2F-associated phosphoprotein |
| PKY56423.1 | K | RNA polymerase II transcription factor SIII (Elongin) subunit A |
| GBC42708.1 | K | COBRA1; negative elongation factor B |
| GBC26955.1 | K | BTB/POZ domain |
| GBC27227.1 | K | BTB/POZ domain |
| PKK76015.1 | K | BTB/POZ domain |
| POG68457.1 | K | BTB/POZ domain |
| POG68473.1 | K | BTB/POZ domain |
| POG71272.1 | K | BTB/POZ domain |
| GBC25627.1 | K | IBR domain, a half RING-finger domain |
| PKC05346.1 | K | Sds3-like |
| PKY12470.1 | K | C2H2-type zinc finger |
| PKK60414.1 | K | FLYWCH zinc finger domain |
| GBC45787.1 | K | Fungal specific transcription factor domain |
| PKK67681.1 | K | GATA zinc finger |
| GBC44099.1 | K | FAR1 DNA-binding domain |
| PKY13633.1 | K | FAR1 DNA-binding domain |
| **Transcription activation/repression** | | |
| POG83212.1 | K | MED23; mediator of RNA polymerase II transcription subunit 23 |
| PKC51537.1 | K | Tetracyclin repressor-like, C-terminal domain |
| **Replication** | | |
| EXX53628.1 | L | INO80; DNA helicase INO80 [EC:3.6.4.12] |
| GBC19234.1 | L | MCM8; DNA helicase MCM8 [EC:3.6.4.12] |
| PKC54699.1 | L | DNA polymerase family B, exonuclease domain |
| EXX79991.1 | L | RAD54L; DNA repair and recombination protein RAD54 and RAD54-like protein [EC:3.6.4.-] |
| EXX78997.1 | L | Initiator Replication protein |
| PKC51926.1 | L | Origin of replication binding protein |
| PKC59544.1 | L | Origin of replication binding protein |
| PKY43752.1 | L | Origin of replication binding protein |
| PKC02080.1 | L | Geminivirus Rep catalytic domain |
| POG63349.1 | L | FEN1; flap endonuclease-1 [EC:3.-.-.-] |
| POG64846.1 | L | POLG; DNA polymerase gamma 1 [EC:2.7.7.7] |
| **Cell wall biogenesis** | | |
| POG65280.1 | M | glmS; glutamine---fructose-6-phosphate transaminase (isomerizing) [EC:2.6.1.16] |
| **Post-translational modification** | | |
| EXX54072.1 | O | NAA35; N-alpha-acetyltransferase 35, NatC auxiliary subunit |
| PKC73274.1 | O | Transglutaminase-like superfamily |
| POG74161.1 | O | PPME1; protein phosphatase methylesterase 1 [EC:3.1.1.89] |
| **Chaperones** | | |
| PKK74043.1 | O | Hsp70 protein |
| EXX78198.1 | O | HSP90A; molecular chaperone HtpG |
| POG75048.1 | O | HSPA1s; heat shock 70kDa protein 1/2/6/8 |
| EXX74441.1 | O | SACS; sacsin |
| PKK75303.1 | O | SACS; sacsin |
| EXX61075.1 | O | CCT2; T-complex protein 1 subunit beta |
| POG71453.1 | O | PFDN1; prefoldin subunit 1 |
| **Protein turnover** | | |
| **Protein ubiquitinylation** | | |
| PKK40735.1 | O | F-box-like |
| POG59781.1 | O | F-box-like |
| EXX75265.1 | O | ATG3; ubiquitin-like-conjugating enzyme ATG3 |
| EXX79052.1 | O | RNF213; E3 ubiquitin-protein ligase RNF213 [EC:2.3.2.27] |

| | | |
|---|---|---|
| GBC23180.1 | O | RNF213; E3 ubiquitin-protein ligase RNF213 [EC:2.3.2.27] |
| POG74668.1 | O | UBE2D; ubiquitin-conjugating enzyme E2 D [EC:2.3.2.23] |
| PKY44117.1 | O | UBE2T; ubiquitin-conjugating enzyme E2 T [EC:2.3.2.23] |
| PKC15495.1 | O | UBLE1A; ubiquitin-like 1-activating enzyme E1 A [EC:6.2.1.45] |
| PKY61708.1 | O | USP7; ubiquitin carboxyl-terminal hydrolase 7 [EC:3.4.19.12] |
| **Proteasome** | | |
| EXX74409.1 | O | PSMA3; 20S proteasome subunit alpha 7 [EC:3.4.25.1] |
| POG79425.1 | O | PSMC4; 26S proteasome regulatory subunit T3 |
| EXX73751.1 | O | PSME4; proteasome activator subunit 4 |
| **Ion transport** | | |
| GBC20706.1 | P | Ion transport protein |
| POG79310.1 | P | NIPA; magnesium transporter |
| **Secondary metabolite metabolism** | | |
| PKY43547.1 | Q | TMLHE; trimethyllysine dioxygenase [EC:1.14.11.8] |
| PKY50609.1 | Q | Cytochrome P450 |
| **Signal transduction** | | |
| POG67284.1 | T | YWHAE; 14-3-3 protein epsilon |
| PKC60368.1 | T | CALM; calmodulin |
| EXX74427.1 | T | DOCK3; dedicator of cytokinesis protein 3 |
| POG68690.1 | T | RRAGC_D; Ras-related GTP-binding protein C/D |
| PKC09489.1 | T | LIM domain |
| EXX71383.1 | T | MAP3K7; mitogen-activated protein kinase kinase kinase 7 [EC:2.7.11.25] |
| PKK68685.1 | T | Protein kinase domain |
| EXX53831.1 | T | Protein tyrosine kinase |
| EXX54173.1 | T | Protein tyrosine kinase |
| EXX54325.1 | T | Protein tyrosine kinase |
| EXX56146.1 | T | Protein tyrosine kinase |
| GBC14767.1 | T | Protein tyrosine kinase |
| GBC17566.1 | T | Protein tyrosine kinase |
| GBC24805.1 | T | Protein tyrosine kinase |
| GBC31826.1 | T | Protein tyrosine kinase |
| GBC40248.1 | T | Protein tyrosine kinase |
| GBC42998.1 | T | Protein tyrosine kinase |
| PKB95537.1 | T | Protein tyrosine kinase |
| PKC10133.1 | T | Protein tyrosine kinase |
| PKY24156.1 | T | Protein tyrosine kinase |
| PKY24378.1 | T | Protein tyrosine kinase |
| PKY30970.1 | T | Protein tyrosine kinase |
| PKY53285.1 | T | Protein tyrosine kinase |
| POG60871.1 | T | Protein tyrosine kinase |
| POG63866.1 | T | Protein tyrosine kinase |
| POG68335.1 | T | Protein tyrosine kinase |
| POG68383.1 | T | Protein tyrosine kinase |
| POG69834.1 | T | Protein tyrosine kinase |
| POG71482.1 | T | Protein tyrosine kinase |
| POG77262.1 | T | Protein tyrosine kinase |
| EXX62365.1 | T | Sel1 repeat |
| GBC22962.1 | T | Sel1 repeat |
| GBC38025.1 | T | Sel1 repeat |
| GBC52916.1 | T | Sel1 repeat |
| PKC05432.1 | T | Sel1 repeat |
| PKY48031.1 | T | Sel1 repeat |
| PKY53316.1 | T | Sel1 repeat |
| POG60589.1 | T | Sel1 repeat |
| GBC18251.1 | T | Tetratricopeptide repeat |
| PKY19547.1 | T | Tetratricopeptide repeat |
| PKY21322.1 | T | Tetratricopeptide repeat |
| POG65388.1 | T | ZAK; sterile alpha motif and leucine zipper containing kinase AZK [EC:2.7.11.25] |
| GBC41753.1 | T | RGD1; Rho GTPase-activating protein RGD1 |
| GBC44208.1 | T | transient-receptor-potential calcium channel |
| PKK80325.1 | T | MAPKAP1; target of rapamycin complex 2 subunit MAPKAP1 |
| PKC70587.1 | T | PLD1_2; phospholipase D1/2 [EC:3.1.4.4] |
| PKY56982.1 | T | Fic/DOC family |
| **Intracellular trafficking** | | |
| **Vesicular transport** | | |
| GBC39457.1 | U | Root hair defective 3 GTP-binding protein (RHD3) |

| | | |
|---|---|---|
| GBC51077.1 | U | ARF1; ADP-ribosylation factor 1 |
| EXX69781.1 | U | RAB6A; Ras-related protein Rab-6A |
| EXX63485.1 | U | RAB2B; Ras-related protein Rab-2B |
| POG68635.1 | U | RAB1A; Ras-related protein Rab-1A |
| PKY21321.1 | U | STX1B_2_3; syntaxin 1B/2/3 |
| **Nuclear transport** | | |
| POG69665.1 | U | KPNB1; importin subunit beta-1 |
| **Endocytosis** | | |
| POG79581.1 | U | PAN1; actin cytoskeleton-regulatory complex protein PAN1 |
| **Defense mechanisms** | | |
| PKK70402.1 | V | ABC transporter transmembrane region |
| PKY44545.1 | V | Restriction endonuclease |
| **Mobilome** | | |
| **Reverse transcriptase** | | |
| GBC21731.1 | X | Reverse transcriptase (RNA-dependent DNA polymerase) |
| GBC25083.1 | X | Reverse transcriptase (RNA-dependent DNA polymerase) |
| GBC43786.1 | X | Reverse transcriptase (RNA-dependent DNA polymerase) |
| GBC46870.1 | X | Reverse transcriptase (RNA-dependent DNA polymerase) |
| **Transposase** | | |
| PKY17692.1 | X | Tc5 transposase DNA-binding domain |
| GBC42558.1 | X | Transposase |
| PKB94031.1 | X | hAT family C-terminal dimerisation region |
| EXX65430.1 | X | MULE transposase domain |
| PKK59383.1 | X | MULE transposase domain |
| PKK63359.1 | X | MULE transposase domain |
| **Cytoskeleton** | | |
| POG72654.1 | Z | ACTB_G1; actin beta/gamma 1 |
| PKY30647.1 | Z | ACTR8; actin-related protein 8 |
| GBC44408.1 | Z | KIF4_21_27; kinesin family member 4/21/27 |
| POG82738.1 | Z | KIF4_21_27; kinesin family member 4/21/27 |
| POG67461.1 | Z | TUBA; tubulin alpha |
| EXX68997.1 | Z | N-acetyltransferase B complex (NatB) non catalytic subunit |
| **General Function prediction only** | | |
| PKY41450.1 | R | ABCF2; ATP-binding cassette, subfamily F, member 2 |
| PKY44157.1 | R | Bromodomain |
| PKC50387.1 | R | 5' nucleotidase, deoxy (Pyrimidine), cytosolic type C protein (NT5C) |
| PKY56279.1 | R | Aldo/keto reductase family |
| PKY61135.1 | R | Eukaryotic aspartyl protease |
| POG73198.1 | R | Glycosyltransferase sugar-binding region containing DXD motif |
| POG73965.1 | R | Membrane bound O-acyl transferase family |
| POG66943.1 | R | Methyltransferase domain |
| GBC41907.1 | R | RNase H |
| PKY51947.1 | R | RNase H |
| GBC51896.1 | R | RNase H-like domain found in reverse transcriptase |
| PKC00510.1 | R | Lycopene cyclase protein |
| GBC28674.1 | R | Zinc knuckle |
| PKY56052.1 | R | Zinc knuckle |
| PKY58643.1 | R | Zinc knuckle |
| GBC40136.1 | R | BED zinc finger |
| PKC53924.1 | R | BED zinc finger |
| **Domains with unknown functions** | | |
| PKK80927.1 | S | NYN domain |
| PKY34399.1 | S | Uncharacterized alpha/beta hydrolase domain (DUF2235) |
| PKK59453.1 | S | TLD |
| GBC20811.1 | S | Protein of unknown function (DUF 659) |
| GBC41695.1 | S | Protein of unknown function (DUF 659) |
| PKY30949.1 | S | Protein of unknown function (DUF 659) |
| PKC17287.1 | S | Protein of unknown function (DUF3684) |
| PKC65729.1 | S | Protein of unknown function (DUF3684) |
| PKY45451.1 | S | Protein of unknown function (Ytp1) |

# CHAPTER III

# GENOMIC ANALYSIS OF THE YET-UNCULTURED BINATOTA REVEALS BROAD METHYLOTROPHIC, ALKANE-DEGRADATION, AND PIGMENT PRODUCTION CAPACITIES

## 3.1 Abstract

The recent leveraging of genome-resolved metagenomics has generated an enormous number of genomes from novel uncultured microbial lineages, yet left many clades undescribed. We here present a global analysis of genomes belonging to the Binatota (UBP10), a globally distributed, yet-uncharacterized bacterial phylum. All orders in the Binatota encoded the capacity for aerobic methylotrophy using methanol, methylamine, sulfomethanes, chloromethanes as substrates. Methylotrophy in the Binatota was characterized by order-specific substrate degradation preferences, as well as extensive metabolic versatility, i.e. the utilization of diverse sets of genes, pathways and combinations to achieve a specific metabolic goal. The genomes also encoded multiple alkane hydroxylases and monooxygenases, potentially enabling growth on a wide range of alkanes and fatty acids. Pigmentation is inferred from a complete pathway for carotenoids (lycopene, $\beta$ and $\gamma$ carotenes, xanthins, chlorobactenes, and spheroidenes) production. Further, the majority of genes involved in bacteriochlorophyll *a*, *c*, and *d* biosynthesis were identified; although absence of key genes and failure to identify a photosynthetic reaction center precludes proposing phototrophic capacities. Analysis of 16S rRNA databases showed Binatota's preferences to terrestrial and freshwater ecosystems, hydrocarbon-rich habitats, and sponges supporting their potential role in mitigating methanol and methane emissions, breakdown of alkanes, as well as their association with sponges. Our results expand the lists of methylotrophic, aerobic alkane degrading, and pigment-producing lineages. We also highlight the consistent encountering of incomplete biosynthetic pathways in microbial genomes, a phenomenon necessitating careful assessment when assigning putative functions based on a set-threshold of pathway completion.

## 3.2 Importance

A wide range of microbial lineages remain uncultured, yet little is known regarding their metabolic capacities, physiological preferences and ecological roles in various ecosystems. We conducted a thorough comparative genomic analysis of 108 genomes belonging to the Binatota (UBP10), a globally distributed, yet-uncharacterized bacterial phylum. We present evidence that members of the order Binatota specialize in methylotrophy, and identify an extensive repertoire of genes and pathways mediating the oxidation of multiple one-carbon (C1) compounds in Binatota genomes. The occurrence of multiple alkane hydroxylases and

monooxygenases in these genomes was also identified, potentially enabling growth on a wide range of alkanes and fatty acids. Pigmentation is inferred from a complete pathway for carotenoids production. We also report on the presence of incomplete chlorophyll biosynthetic pathways in all genomes, and propose several evolutionary-grounded scenarios that could explain such pattern. Assessment of the ecological distribution patterns of the Binatota indicates preference of its members to terrestrial and freshwater ecosystems characterized by high methane and methanol emissions, as well multiple hydrocarbon-rich habitats, and marine sponges.

## 3.3 Introduction

Approaches that directly recover genomes from environmental samples, e.g. single-cell genomics and genome-resolved metagenomics, and hence bypass the hurdle of cultivation have come of age in the last decade. The resulting availability of environmentally-sourced genomes, obtained as SAGs (single amplified genomes) or MAGs (metagenome-assembled genomes) is having a lasting impact on the field of microbial ecology. Distinct strategies are employed for the analysis of the deluge of obtained genomes. Site- or habitat-specific studies focus on spatiotemporal sampling of a single site or habitat of interest. Function-based studies focus on genomes from single or multiple habitats to identify and characterize organisms involved in a specific process, e.g. cellulose degradation [104] or sulfate-reduction [9]. Phylogeny-oriented (phylocentric) studies, on the other hand, focus on characterizing genomes belonging to a specific lineage of interest. The aim of such studies is to delineate pan, core, and dispensable gene repertoires for a target lineage, document its defining metabolic capabilities [24, 328], understand its putative roles in various habitats [119, 169], and elucidate genomic basis underpinning the observed niche specializing patterns [419]. The scope of phylocentric studies could range from the analysis of a single genome from a single ecosystem [411], to global sampling and *in-silico* analysis efforts [23, 327]. The feasibility and value of phylocentric strategies have recently been enhanced by the development of a genome-based (phylogenomic) taxonomic outline based on extractable data from MAGs and SAGs providing a solid framework for knowledge building and data communication [293], as well as recent efforts for massive, high-throughput binning of genomes from global collections of publicly available metagenomes in GenBank nr and Integrated Microbial Genomes & Microbiomes (IMG/M) database [280, 298]. As such, these studies provide immensely useful information on potential metabolic capabilities and physiological preferences of yet-uncultured taxa. However, such in-silico predictions require confirmation through enrichment, isolation, complementary cloning and expression studies of gene of interest, or other functional genomics approaches to ascertain their phenotypic relevance.

Candidate phylum UBP10 has originally been described as one of the novel lineages recovered from a massive binning effort that reconstructed thousands of genomes from publicly available metagenomic datasets [295]. UBP10 has subsequently been named candidate phylum Binatota (henceforth Binatota) in an effort to promote nomenclature for uncultured lineages based on attributes identified in MAGs and SAGs [77]. The recent generation of 52,515 distinct MAGs binned from over 10,000 metagenomes [280] has greatly increased the number of available Binatota genomes. Here, we utilize a phylocentric approach and present a comparative analysis of the putative metabolic and biosynthetic capacities and putative

ecological roles of members of the candidate phylum Binatota, as based on sequence data from 108 MAGs. Our study documents aerobic methylotrophy, aerobic alkane degradation, and carotenoid pigmentation as defining traits in the Binatota. We also highlight the presence of incomplete chlorophyll biosynthetic pathways in all genomes, and propose several evolutionary-grounded scenarios that could explain such pattern.

## 3.4 Results

### 3.4.1 Genomes analyzed in this study

A total of 108 Binatota MAGs with >70% completion and <10% contamination were used for this study, which included 86 medium-quality (>50% completion, <10% contamination) and 22 high-quality (>90% completion, <5% contamination) genomes, as defined by MIMAG standards [37]. Binatota genomes clustered into seven orders designated as Bin18 (n=2), Binatales (n=48), HRBin30 (n=7), UBA1149 (n=9), UBA9968 (n=34), UBA12015 (n=1), UTPRO1 (n=7), encompassing 12 families, and 24 genera (Figure 1, Table S1). 16S rRNA gene sequences extracted from orders Bin18 and UBA9968 genomes were classified in SILVA (release 138) [315] as members of class bacteriap25 in the phylum Myxococcota, order Binatales and order HRBin30 as uncultured phylum RCP2-54, and orders UBA1149 and UTPRO1 as uncultured Desulfobacterota classes (Table S1). RDP II-classification (July 2017 release, accessed July 2020) classified all Binatota sequences as unclassified Deltaproteobacteria (Table S1).

### 3.4.2 Methylotrophy in the Binatota

#### Methanol

With the exception of HRBin30, all orders encoded at least one type of methanol dehydrogenase (Figure 2a). Three distinct types of methanol dehydrogenases were identified (Figure 2a, b): (1.) the NAD(P)-binding MDO/MNO-type methanol dehydrogenase (*mno*), typically associated with Gram-positive methylotrophic bacteria (Actinobacteria and *Bacillus methanolicus*) [160], was the only type of methanol dehydrogenase identified in orders UBA9968, UBA12105, and UTPRO1 (Figure 2a, Extended data 1), as well as some UBA1149 and Binatales genomes. (2.) the MDH2-type methanol dehydrogenase, previously discovered in members of the Burkholderiales and Rhodocyclales [193], was encountered in the majority of order UBA1149 genomes and in two Binatales genomes. (3.) the lanthanide-dependent pyrroloquinoline quinone (PQQ) methanol dehydrogenase XoxF-type was encountered in nine genomes from the orders Bin18, and Binatales, together with the accessory XoxG c-type cytochrome and XoxJ periplasmic-binding proteins (Figure 2a). All later genomes also encoded PQQ biosynthesis. Surprisingly, none of the genomes encoded the MxaF1-type (MDH1) methanol dehydrogenase, typically encountered in model methylotrophs [72].

#### Methylamine

All Binatota orders except UBA9968 encoded methylamine degradation capacity. The direct periplasmic route (methylamine dehydrogenase; *mau*) was more common, with *mauA* and

*mauB* enzyme subunits encoded in the Binatales, HRBin30, UBA1149, UBA12105, and UTPRO1 (Figure 2a, Extended data 1). Amicyanin (encoded by *mauC*) is the most probable electron acceptor for methylamine dehydrogenase [72] (Figure 2a). On the other hand, one Bin18 genome, and two Binatales genomes (that also encode the *mau* cluster) encoded the full complement of genes for methylamine oxidation via the indirect glutamate pathway (Figure 2a, Extended data 1).

**Methylated sulfur compounds**

Binatota genomes encoded several enzymes involved in the degradation of dimethyl sulfone, methane sulfonic acid (MSA), and dimethyl sulfide (DMS). Nine genomes (two Bin18, and 7 Binatales) encoded dimethyl sulfone monooxygenase (*sfnG*) involved in the degradation of dimethyl sulfone to MSA with the concomitant release of formaldehyde. Three of these nine genomes also encoded alkane sulfonic acid monooxygenase (*ssuD*), which will further degrade the MSA to formaldehyde and sulfite. Degradation of DMS via DMS monooxygenase (*dmoA*) to formaldehyde and sulfide was encountered in 13 genomes (2 Bin18, 9 Binatales, and 2 UBA9968). Further, one Binatales genome encoded the *dso* system (EC: 1.14.13.245) for DMS oxidation to dimethyl sulfone, which could be further degraded to MSA as explained above (Figure 2a, Extended data 1).

**Dihalogenated methane**

One Bin18 genome encoded the specific dehalogenase/ glutathione S-transferase (*dcmA*) capable of converting dichloromethane to formaldehyde.

**Methane**

Genes encoding copper membrane monooxygenases (CuMMOs), a family of enzymes that includes particulate methane monooxygenase (pMMO) were identified in orders Bin18 (2/2 genomes) and Binatales (9/48 genomes) (Figure 2a, Extended data 1), while genes encoding soluble methane monooxygenase (sMMO) were not found. A single copy of the three genes encoding all CuMMO subunits (A, B, and C) was encountered in 9 of the 11 genomes, while two copies were identified in two genomes. CuMMO subunit-encoding genes (A, B, and C) occurred as a contiguous unit in all genomes, with a CAB (5 genomes), and/or CAxB or CAxxB (8 genomes, where x is a hypothetical protein) organization, similar to the pMMO operon structure in methanotrophic Proteobacteria, Verrucomicrobia, and *Candidatus* Methylomirabilis (NC10) [116, 117, 180, 325] (Figure 2c). In addition, five of the above-mentioned eleven genomes also encoded a *pmoD* subunit, recently suggested to be involved in facilitating the enzyme complex assembly, and/or in electron transfer to the enzyme's active site [125, 332]. Phylogenetic analysis of Binatota *pmoA* sequences revealed their affiliation with two distinct clades: the yet-uncultured Cluster 2 TUSC (Tropical Upland Soil Cluster) methanotrophs [205] (2 Binatales genomes), and a clade encompassing *bmoA* sequences (putative butane monooxygenase gene A) from Actinobacteria (*Nocardioides* sp. strain CF8, *Mycolicibacterium*, and *Rhodococcus*) and SAR324 (*Candidatus* Lambdaproteobacteria) [234, 348] (Figure 2d). Members harboring these specific lineages have previously been identified in a wide range of environments, including soil [205]. Previous studies

have linked Cluster 2 TUSC CuMMO-harboring organisms to methane oxidation based on selective enrichment on methane in microcosms derived from Lake Washington sediments [194]. Binatota genomes encoding TUSC-affiliated CuMMO also encoded genes for downstream methanol and formaldehyde oxidation as well as formaldehyde assimilation (see below), providing further evidence for their putative involvement in methane oxidation. On the other hand, studies on *Nocardioides* sp. strain CF8 demonstrated its capacity to oxidize short chain (C2-C4) hydrocarbons, but not methane, via its CuMMO, and its genome lacked methanol dehydrogenase homologues [149]. Such data favor a putative short chain hydrocarbon degradation function for organisms encoding this type of CuMMO, although we note that five out of the nine Binatota genomes encoding SAR324/ Actinobacteria-affiliated *pmoA* sequences also encoded at least one methanol dehydrogenase homologue. Modeling CuMMO subunits from both TUSC-type and Actinobacteria/SAR324-type Binatota genomes using *Methylococcus capsulatus* (Bath) 3D model (Protein DataBank ID: 3rbg) revealed a heterotrimeric structure ($\alpha_3\beta_3\gamma_3$) with the 7, 2, and 5 alpha helices of the PmoA, PmoB, and PmoC subunits, respectively, as well as the beta sheets characteristic of PmoA, and PmoB subunits (Figure S1). Recently, the location of the active site at the amino terminus of the PmoB subunit has been suggested [19]. There has been recent debate as to the exact nuclearity of the Cu cofactor at the active site [19, 57, 336]. Regardless of the nuclearity of the copper metal center, conserved histidine residues $His^{33}$, $His^{137}$, and $His^{139}$ (numbering following the *Methylococcus capsulatus* str. Bath PmoB subunit, Protein DataBank ID: 3rbg), thought to coordinate the Cu cofactor, were identified in all TUSC-affiliated and SAR324/ Actinobacteria-affiliated Binatota CuMMO sequences (Figure S1). Modeling PmoB subunits from both TUSC-type and Actinobacteria/SAR324-type Binatota genomes using *Methylococcus capsulatus* (Bath) PmoB subunit (PDB ID: 3rbg) predicted the binding pockets for Cu in Binatota sequences (Figure 2e).

As previously noted [72], methylotrophy requires the possession of three metabolic modules for: C1 oxidation to formaldehyde, formaldehyde oxidation to $CO_2$, and formaldehyde assimilation. Formaldehyde generated by C1 substrates oxidation is subsequently oxidized to formate and eventually $CO_2$. Multiple pathways for formaldehyde oxidation to formate were identified in all Binatota orders (Supp. text, Figure S2). In addition, the majority of Binatota genomes encoded a formate dehydrogenase for formate oxidation to $CO_2$ (Supp. text, Figure S2). Finally, for assimilating formaldehyde into biomass, genes encoding all enzymes of the serine cycle were identified in all genomes, as well as genes encoding different routes of glyoxylate regeneration (Supp. text, Figure S2).

### 3.4.3 Alkane degradation in the Binatota

Beside methylotrophy and methanotrophy, Binatota genomes exhibited extensive short-, medium-, and long-chain alkanes degradation capabilities. In addition to the putative capacity of Actinobacteria/SAR324-affiliated CuMMO to oxidize $C_1$-$C_5$ alkanes, and $C_1$-$C_4$ alkenes as described above, some Binatota genomes encoded propane-2-monoxygenase (*prmABC*), an enzyme mediating propane hydroxylation in the 2-position yielding isopropanol. Several genomes, also encoded medium chain-specific alkane hydroxylases, e.g. homologues of the nonheme iron *alkB* [69] and Cyp153-class alkane hydroxylases [386]. The genomes also encoded multiple long-chain specific alkane monooxygenase, e.g. *ladA* homo-

logues (EC:1.14.14.28) [233] (Figure 3, Extended data 1). Finally, Binatota genomes encoded the capacity to metabolize medium-chain haloalkane substrates. All genomes encoded *dhaA* (haloalkane dehalogenases [EC:3.8.1.5]) known to have a broad substrate specificity for medium chain length (C3 to C10) mono-, and dihaloalkanes, resulting in the production of their corresponding primary alcohol, and haloalcohols, respectively [276] (Figure 3, Extended data 1).

Alcohol and aldehyde dehydrogenases sequentially oxidize the resulting alcohols to their corresponding fatty acids or fatty acyl-CoA. Binatota genomes encode a plethora of alcohol and aldehyde dehydrogenases mediating such processes (supp. text, Figure S3). As well, a complete fatty acid degradation machinery that enables all orders of the Binatota to degrade short-, medium-, and long-chain fatty acids to acetyl CoA and propionyl-CoA were identified (supp. text, Figure S3).

### 3.4.4 Predicted electron transport chain

All Binatota genomes encode an aerobic respiratory chain comprising complexes I, II, alternate complex III (ACIII, encoded by *actABCDEFG*), and complex IV, as well as an F-type $H^+$-translocating ATP synthase (Supp. text, Figure 4). Binatota genomes also encode respiratory $O_2$-tolerant $H_2$-uptake [NiFe] hydrogenases, belonging to groups 1c (6 sequences), 1f (22 sequences), 1i (1 sequence), and 1h (4 sequences) (Figure S4). Simultaneous oxidation of hydrogen (via type I respiratory $O_2$-tolerant hydrogenases) and methane (via pMMO) has been shown to occur in methanotrophic Verrucomicrobia to maximize proton-motive force generation and subsequent ATP production [61]. As well, some of the reduced quinones generated through $H_2$ oxidation are thought to provide reducing power for catalysis by pMMO [61] (Figure 4). Details on the distribution of ETC components across Binatota orders are shown in Figure S4, and the proposed electron flow under different growth conditions are presented in the supplementary text.

### 3.4.5 Pigment production genes in the Binatota

#### Carotenoids

Analysis of the Binatota genomes demonstrated a wide range of hydrocarbon (carotenes) and oxygenated (xanthophyll) carotenoid biosynthesis capabilities. Carotenoids biosynthetic machinery in the Binatota included *crtB* for 15-cis-phyotene synthesis from geranylgeranyl-PP; *crtI*, *crtP*, *crtQ*, and *crtH* for neurosporene and all-*trans* lycopene formation from 15-cis-phytone; *crtY* or *crtL* for gamma- and beta-carotene formation from all-*trans* lycopene; and a wide range of genes encoding enzymes for the conversion of neurosporene to spheroidene and 7,8-dihydro $\beta$-carotene, as well as the conversion of all-trans lycopene to spirilloxanthin, gamma-carotene to hydroxy-chlorobactene glucoside ester and hydroxy-$\gamma$-carotene glucoside ester, and beta carotene to isorenieratene and zeaxanthins (Figures 5a-b, Extended data 1). Gene distribution pattern (Figure 5a, Extended data 1) predicts that all Binatota orders are capable of neurosporene and all-trans lycopene biosynthesis, and all but the order HRBin30 are capable of isorenieratene, zeaxanthin, $\beta$-carotene and dihydro $\beta$-carotene biosynthesis, and with specialization of order UTPRO1 in spirilloxanthin, spheroidene, hydroxy-chlorobactene, and hydroxy $\gamma$-carotene biosynthesis.

**Bacteriochlorophylls**

Surprisingly, homologues of multiple genes involved in bacteriochlorophyll biosynthesis were ubiquitous in Binatota genomes (Figure 6a-c). Bacteriochlorophyll biosynthesis starts with the formation of chlorophyllide *a* from protoporphyrin IX (Figure 6b). Within this pathway, genes encoding the first *bchI* (Mg-chelatase [EC:6.6.1.1]), third *bchE* (magnesium-protoporphyrin IX monomethyl ester cyclase [EC:1.21.98.3]), and fourth *bchLNB* (3,8-divinyl protochlorophyllide reductase [EC:1.3.7.7]) steps were identified in the Binatota genomes (Figures 6a, 6b, Extended data 1). However, homologues of genes encoding the second *bchM* (magnesium-protoporphyrin O-methyltransferase [EC:2.1.1.11]), and the fifth (*bciA* or *bicB* (3,8-divinyl protochlorophyllide *a* 8-vinyl-reductase), or *bchXYZ* (chlorophyllide *a* reductase, EC 1.3.7.15])) steps were absent (Figure 6a-b). A similar patchy distribution was observed in the pathway for bacteriochlorophyll *a* (Bchl *a*) formation from chlorophyllide *a* (Figure 6b), where genes encoding *bchXYZ* (chlorophyllide *a* reductase [EC 1.3.7.15]) and *bchF* (chlorophyllide *a* $3^1$-hydratase [EC 4.2.1.165]) were not identified, while genes encoding *bchC* (bacteriochlorophyllide *a* dehydrogenase [EC 1.1.1.396]), *bchG* (bacteriochlorophyll a synthase [EC:2.5.1.133]), and *bchP* (geranylgeranyl-bacteriochlorophyllide *a* reductase [EC 1.3.1.111)) were present in most genomes (Figure 6a, Extended data 1). Finally, within the pathway for bacteriochlorophylls *c* (Bchl *c*) and *d* (Bchl *d*) formation from chlorophyllide *a* (Figure 6b), genes for *bciC* (chlorophyllide a hydrolase [EC:3.1.1.100]), and *bchF* (chlorophyllide *a* $3^1$-hydratase [EC:4.2.1.165]) or *bchV* (3-vinyl bacteriochlorophyllide hydratase [EC:4.2.1.169] were not identified, while genes for *bchR* (bacteriochlorophyllide d C-12(1)-methyltransferase [EC:2.1.1.331]), *bchQ* (bacteriochlorophyllide d C-8(2)-methyltransferase [EC:2.1.1.332]), *bchU* (bacteriochlorophyllide d C-20 methyltransferase [EC:2.1.1.333]), and *bchK* (bacteriochlorophyll c synthase [EC:2.5.1.-]) were identified (Figure 6b, Extended data 1).

### 3.4.6 Ecological distribution of the Binatota

A total of 1,889 (GenBank nt) and 1,213 (IMG/M) 16S rRNA genes affiliated with the Binatota orders were identified (Extended data 2, Figures 7, S5a). Analyzing their environmental distribution showed preference of Binatota to terrestrial soil habitats (39.5-83.0% of GenBank, 31.7-91.6% of IMG/M 16S rRNA gene sequences in various orders), as well as plant-associated (particularly rhizosphere) environments; although this could partly be attributed to sampling bias of these globally distributed and immensely important ecosystems (Figure 7a). On the other hand, a paucity of Binatota-affiliated sequences was observed in marine settings, with sequences absent or minimally present for Binatales, HRBin30, UBA9968, and UTPRO1 datasets (Figure 7a). The majority of sequences from marine origin were sediment-associated, being encountered in hydrothermal vents, deep marine sediments, and coastal sediments, with only the Bin18 sequences sampled from IMG/M showing representation in the vast, relatively well-sampled pelagic waters (Figure 7d).

In addition to the 16S rRNA-based analysis, we queried the datasets from which a Binatota MAG was binned using the sequence of their ribosomal protein S3, and estimating the Binatota relative abundance as the number of reads mapped to contigs with a Binatota ribosomal protein S3 as a percentage of the number of reads mapped to all contigs encoding a

ribosomal protein S3 gene. Results showed relative abundances ranging between 0.1-10.21% (average 3.84±3.21 %) (Table S1).

In addition to phylum-wide patterns, order-specific environmental preferences were also observed. For example, in order Bin18, one of the two available genomes originated from the Mediterranean sponge *Aplysina aerophoba*. Analysis of the 16S rRNA dataset suggests a notable association between Bin18 and sponges, with a relatively high host-associated sequences (Figure 7a), the majority of which (58.3% NCBI-nt, 25.0% IMG/M) were recovered from the Porifera microbiome (Figures 7e, S5f). Bin18-affiliated 16S rRNA gene sequences were identified in a wide range of sponges from ten genera and five global habitat ranges (the Mediterranean genera *Ircinia*, *Petrosia*, *Chondrosia*, and *Aplysina*, the Caribbean genera *Agelas*, *Xestospongia*, and *Aaptos*, the Indo-West Pacific genus *Theonella*, the Pacific Dysideidae family, and the Great Barrier Reef genus *Rhopaloeides*), suggesting its widespread distribution beyond a single sponge species. The absolute majority of order Binatales sequences (83.0% NCBI-nt, 91.6% IMG/M) were of a terrestrial origin (Figures 7a, S5c), in addition to multiple rhizosphere-associated samples (7.5% NCBI-nt and 2.8% IMG/M, respectively) (Figure 7a, S5f). Notably, a relatively large proportion of Binatales soil sequences originated either from wetlands (peats, bogs) or forest soils (Figures 7b, S5c), strongly suggesting the preference of the order Binatales to acidic and organic/methane-rich terrestrial habitats. This corresponds with the fact that 42 out of 48 Binatales genomes were recovered from soil, 38 of which were from acidic wetland or forest soils (Figure 1, Table S1). Genomes of UBA9968 were recovered from a wide range of terrestrial and non-marine aquatic environments, and the observed 16S rRNA gene distribution verifies their ubiquity in all but marine habitats (Figures 7a, S5b-g). Finally, while genomes from orders HRBin30, UBA1149 and UTPRO1 were recovered from limited environmental settings (thermal springs for HRBin30, gaseous hydrocarbon impacted habitats, e.g. marine hydrothermal vents and gas-saturated Lake Kivu for UBA1149, and soil and hydrothermal environments for UTPRO1) (Figure 1, Table S1), 16S rRNA gene analysis suggested their presence in a wide range of environments from each macro-scale environment classification (Figures 7a, S5b-g).

## 3.5 Discussion

### 3.5.1 Expanding the world of methylotrophy

The current study expands the list of lineages potentially capable of methylotrophy. An extensive repertoire of genes and pathways mediating the oxidation of multiple C1 compounds to formaldehyde (Figure 2, 4), formaldehyde oxidation to $CO_2$ (Figure S2), as well as formaldehyde assimilation pathways (Figure S2) were identified, indicating that such capacity is a defining metabolic trait in the Binatota. A certain degree of order-level substrate preference was observed, with potential utilization of methanol in all orders except HRBin30, methylamine in all orders except UBA9968, S-containing C1 compound in Bin18, Binatales, and UBA9968, halogenated methane in Bin18, and possible methane utilization (methanotrophy) in Bin18 and Binatales (Figure 2a).

Aerobic methylotrophy has been documented in members of the alpha, beta, and gamma Proteobacteria [74], Bacteroidetes [32], Actinobacteria (e.g. genera *Arthrobacter* and *Mycobacterium*), Firmicutes (e.g. *Bacillus methanolicus*) [251], Verrucomicrobia [307], and

*Candidatus* Methylomirabilis (NC10) [118]. Further, studies employing genome-resolved metagenomics identified some signatures of methylotrophy, e.g. methanol oxidation [97, 419], formaldehyde oxidation/ assimilation [48], and methylamine oxidation [419], in the Gemmatimonadetes, Rokubacteria, Chloroflexi, Actinobacteria, Acidobacteria, and Lambdaproteobacteria. The possible contribution of Binatota to methane oxidation (methanotrophy) is especially notable, given the global magnitude of methane emissions, and the relatively narrower range of organisms (Proteobacteria, Verrucomicrobia, and *Candidatus* Methylomirabilota (NC10)) [203] capable of this special type of methylotrophy. As described above, indirect evidence exists for the involvement of Binatota harboring TUSC-type CuMMO sequences in methane oxidation, while it is currently uncertain whether Binatota harboring SAR324/Actinobacteria-type CuMMO sequences are involved in oxidation of methane, gaseous alkanes, or both. pMMO of methanotrophs is also capable of oxidizing ammonia to hydroxylamine, which necessitates methanotrophs to employ hydroxylamine detoxification mechanisms [261]. All eleven Binatota genomes encoding CuMMO also encoded at least one homologue of *nir*, *nor*, and/or *nos* genes that could potentially convert harmful N-oxide byproducts to dinitrogen.

As previously noted [72], methylotrophy requires the possession of three metabolic modules: C1 oxidation to formaldehyde, formaldehyde oxidation to $CO_2$, and formaldehyde assimilation. Within the world of methylotrophs, a wide array of functionally redundant enzymes and pathways has been characterized that mediates various reactions and transformations in such modules. In addition, multiple combinations of different modules have been observed in methylotrophs, with significant variations existing even in phylogenetically related organisms. Our analysis demonstrates that such metabolic versatility indeed occurs within Binatota's methylotrophic modules. While few phylum-wide characteristics emerged, e.g. utilization of serine pathway for formaldehyde assimilation, absence of $H_4$MPT-linked formaldehyde oxidation, and potential utilization of PEP carboxykinase (*pckA*) rather than PEP carboxylase (*ppc*) for $CO_2$ entry to the serine cycle, multiple order-specific differences were observed, e.g. XoxF-type methanol dehydrogenase encoded by Bin18 and Binatales genomes, MDH2-type methanol dehydrogenase encoded by UBA1149 genomes, absence of methanol dehydrogenase homologues in HRBin30 genomes, absence of methylamine oxidation in order UBA9968, and potential utilization of the ethylmalonyl-CoA pathway for glyoxylate regeneration by the majority of the orders versus the glyoxylate shunt by UBA9968.

### 3.5.2 Alkane degradation in the Binatota

A second defining feature of the phylum Binatota, besides methylotrophy, is the widespread capacity for aerobic alkane degradation, as evident by the extensive arsenal of genes mediating aerobic degradation of short- (*prmABC*, propane monooxygenase), medium- (*alkB*, cyp153), and long-chain alkanes (*ladA*) identified (Figure 3), in addition to complete pathways for odd- and even-numbered fatty acids oxidation (Figure S3). Hydrocarbons, including alkanes, have been an integral part of the earth biosphere for eons, and a fraction of microorganisms has evolved specific mechanisms ($O_2$-dependent hydroxylases and monooxygenases, anaerobic addition of fumarate) for their activation and conversion to central metabolites [310]. Aerobic alkane degradation capacity has so far been encountered in the Actinobacteria, Proteobacteria, Firmicutes, Bacteroidetes, as well as in a few Cyanobacteria [310]. As

such, this study adds to the expanding list of phyla capable of aerobic alkane degradation.

### 3.5.3 Metabolic traits explaining niche preferences in the Binatota

Analysis of 16S rRNA gene datasets indicated that the Binatota display phylum-wide (preference to terrestrial habitats and methane/hydrocarbon-impacted habitats, and rarity in pelagic marine environments), as well as order-specific (Bin18 in sponges, HRBin30 and UBA1149 in geothermal settings, Binatales in peats, bogs, and forest soils) habitat preferences (Figures 7, S5). Such distribution patterns could best be understood in light of the phylum's predicted metabolic capabilities. Soils represent an important source of methane, generated through microoxic and anoxic niches within soil's complex architecture [226]. Methane emission from soil is especially prevalent in peatlands, bogs, and wetlands, where incomplete aeration and net carbon deposition occurs. Indeed, anaerobic [394], fluctuating [158], and even oxic [13] wetlands represent one of the largest sources of methane emissions to the atmosphere. As well, terrestrial ecosystems represent a major source of global methanol emissions [207], with its release mostly mediated by demethylation reactions associated with pectin and other plant polysaccharides degradation. C1-metabolizing microorganisms significantly mitigate methane and methanol release to the atmosphere from terrestrial ecosystems [80], and we posit that members of the Binatota identified in soils, rhizosphere, and wetlands contribute to such process. The special preference of order Binatales to acidic peats, bogs, forests, and wetlands could reflect a moderate acidophilic specialization for this order and suggest their contribution to the process in these habitats.

Within the phylum Binatota, it appears that orders HRBin30 and UBA1149 are abundant in thermal vents, thermal springs, and thermal soils, suggesting a specialization to high temperature habitats (Figure 7). Binatota's presence in such habitats could be attributed to high concentrations of alkanes typically encountered in such habitats. Hydrothermal vents display steep gradients of oxygen in their vicinity, emission of high levels of methane and other gaseous alkanes, as well as thermogenic generation of medium-and long-chain alkanes [250]. Indeed, the presence and activity of aerobic hydrocarbon degraders in the vicinity of hydrothermal vents have been well established [234, 348, 392].

The recovery of Binatota genomes from certain lakes could be a reflection of the high gaseous load in such lakes. Multiple genomes, and a large number of Binatota-affiliated 16S rRNA sequences were binned, and identified from Lake Kivu, a meromictic lake characterized by unusually high concentrations of methane [300]. Biotically, methane evolving from Lake Kivu is primarily oxidized by aerobic methanotrophs in surface waters [35, 241, 300], and members of the Binatota could contribute to this process. Binatota genomes were also recovered from Lake Washington sediments, a location that has long served as a model for studying methylotrophy [71, 75]. Steep counter gradients of methane and oxygen occurring in the Lake's sediments enable aerobic methanotrophy to play a major role in controlling methane flux through the water column [15, 16, 73, 220].

Finally, the occurrence and apparent wide distribution of members of the Binatota in sponges, particularly order Bin18, is notable, and could possibly be viewed in terms of the wider symbiotic relationship between sponges and their microbiome. Presence of hydrocarbon-degraders [14, 382], including methanotrophs [338] in the sponge microbiome has previously been noted, especially in deep-water sponges, where low levels of planktonic

biomass restrict the amount of food readily acquired via filter feeding and hence biomass acquisition via methane and alkane oxidation is especially valuable.

### 3.5.4 Carotenoid pigmentation: occurrence and significance

The third defining feature of the Binatota, in addition to aerobic methylotrophy and alkane degradation, is the predicted capacity for carotenoid production. In photosynthetic organisms, carotenoids increase the efficiency of photosynthesis by absorbing in the blue-green region then transferring the absorbed energy to the light-harvesting pigments [153]. Carotenoid production also occurs in a wide range of non-photosynthetic bacteria belonging to the Alpha-, Beta, and Gamma-Proteobacteria (including methano- and methylotrophs), Bacteroidetes, Deinococcus, Thermus, Delta-Proteobacteria, Firmicutes, Actinobacteria, Planctomycetes, and Archaea, e.g. Halobacteriaceae, and *Sulfolobus*. Here, carotenoids could serve as antioxidants [121], and aid in radiation, UV, and desiccation resistance [107, 214]. The link between carotenoid pigmentation and methylo/methanotrophy has long been observed [39], with the majority of known model Alpha- and Gamma-Proteobacteria methano- and methylotrophs being carotenoid producers, although several Gram-positive methylotrophs (*Mycobacterium*, *Arthrobacter*, and *Bacillus*) are not pigmented. Indeed, root-associated facultative methylotrophs of the genus *Methylobacterium* have traditionally been referred to as "pink pigmented facultative methylotrophs" and are seen as integral part of root ecosystems [183]. The exact reason for this correlation is currently unclear and could be related to the soil environment where they are prevalent, where periodic dryness and desiccation could occur, or to the continuous exposure of these aerobes in some habitats to light (e.g. in shallow sediments), necessitating protection from UV exposure.

### 3.5.5 Chlorophyll biosynthesis genes in the Binatota

Perhaps the most intriguing finding in this study is the identification of the majority of genes required for the biosynthesis of bacteriochlorophylls from protoporphyrin-IX (six out of ten genes for bacteriochlorophyll *a* and seven out of eleven genes for bacteriochlorophyll *c* and *d*). While such pattern is tempting to propose phototrophic capacities in the Binatota, the consistent absence of critical genes (*bchM* methyltransferase, *bciA*/*bciB*/*bchXYZ* reductases, *bciC* hydrolase, and *bchF*/*V* hydratases), coupled with our inability to detect reaction center-encoding genes, prevents such a proclamation. Identification of a single or few gene shrapnel from the chlorophyll biosynthesis pathway in microbial genomes is not unique. Indeed, searching the functionally annotated bacterial tree of life AnnoTree [253] using single KEGG orthologies implicated in chlorophyll biosynthesis identifies multiple (in some cases thousands) hits in genomes from non-photosynthetic organisms (Figure 6c). This is consistent with the identification of a *bchG* gene in a Bathyarchaeota fosmid clone [254], and, more recently, a few bacteriochlorophyll synthesis genes in an Asgard genome [239]. However, it should be noted that the high proportion of genes in the bacteriochlorophyll biosynthetic pathway identified in the Binatota genomes has never previously been encountered in non-photosynthetic microbial genomes. Indeed, a search in AnnoTree for the combined occurrence of all seven bacteriochlorophyll synthesis genes identified in Binatota genomes yielded only photosynthetic organisms.

Accordingly, we put forward three scenarios to explain the proposed relationship between Binatota and phototrophy. The most plausible scenario, in our opinion, is that members of the Binatota are pigmented non-photosynthetic organisms capable of carotenoid production, but incapable of chlorophyll production and lack a photosynthetic reaction center. The second scenario posits that members of the Binatota are indeed phototrophs, possessing a complete pathway for chlorophyll biosynthesis and a novel type of reaction center that is bioinformatically unrecognizable. A minimal photosynthetic electron transport chain, similar to *Chloroflexus aurantiacus* [129], with the yet-unidentified reaction center, quinone, alternate complex III (or complex III) and some type of cytochrome c would possibly be functional. Under such scenario, members of the Binatota would be an extremely versatile photoheterotrophic facultative methylotrophic lineage. While such versatility, especially coupling methylotrophy to phototrophy, is rare [76], it has previously been observed in some Rhodospirillaceae species [316]. A third scenario is that Binatota are capable of chlorophyll production, but still incapable of conducting photosynthesis. Under this scenario, genes missed in the pathway are due to shortcomings associated with *in-silico* prediction and conservative gene annotation. For example, the missing *bchM* (E.C.2.1.1.11) could possibly be encoded for by general methyltransferases (EC: 2.1.1.-), the missing *bciC* (EC:3.1.1.100) could possibly be encoded for by general hydrolases (EC: 3.1.1.-), while the missing *bchF* (EC:4.2.1.165) or *bchV* (EC:4.2.1.169) could possibly be encoded for by general hydratases (EC: 4.2.1.-).

Encountering incomplete pathways in genomes of uncultured lineages is an exceedingly common occurrence in SAG and MAG analysis [10, 174]. In many cases, this could plausibly indicate an incomplete contribution to a specific biogeochemical process, e.g. incomplete denitrification of nitrate to nitrite but not ammonia [174], or reduction of sulfite, but not sulfate, to sulfide [79], provided the thermodynamic feasibility of the proposed partial pathway, and, preferably, prior precedence in pure cultures. In other cases, a pattern of absence of peripheral steps could demonstrate the capability for synthesis of a common precursor, e.g., synthesis of precorrin-2 from uroporphyrinogen, but lack of the peripheral pathway for corrin ring biosynthesis leading to an auxotrophy for vitamin B12. Such auxotrophies are common in the microbial world and could be alleviated by nutrient uptake from the outside environment [134] or engagement in a symbiotic lifestyle [88]. However, arguments for metabolic interdependencies, syntrophy, or auxotrophy could not be invoked to explain the consistent absence of specific genes in a dedicated pathway, such as bacteriochlorophyll biosynthesis, especially when analyzing a large number of genomes from multiple habitats. As such, we here raise awareness that using a certain occurrence threshold to judge a pathway's putative functionality could lead to misinterpretations of organismal metabolic capacities due to the frequent occurrence of partial, non-functional, pathways and "gene shrapnel" in microbial genomes.

### 3.5.6 Summary

In conclusion, our work provides a comprehensive assessment of the yet-uncultured phylum Binatota, and highlights its aerobic methylotrophic and alkane degradation capacities, as well as its carotenoid production, and abundance of bacteriochlorophyll synthesis genes in its genomes. Future efforts should focus on confirming these *in-silico* predicted capabilities

and characteristics through targeted enrichment and isolation efforts, as well as functional genomics approaches. We also propose a role for this lineage in mitigating methanol and perhaps even methane emissions from terrestrial and freshwater ecosystems, alkanes degradation in hydrocarbon-rich habitats, and nutritional symbiosis with marine sponges. We present specific scenarios that could explain the unique pattern of chlorophyll biosynthesis gene occurrence, and stress the importance of detailed analysis of pathways completion patterns for appropriate functional assignments in genomes of uncultured taxa.

## 3.6 Materials and Methods

### 3.6.1 Genomes

All genomes classified as belonging to the Binatota in the GTDB database (n=22 MAGs, April 2020) were downloaded as assemblies from NCBI. In addition, 128 metagenome-assembled genomes with the classification "Bacteria;UBP10" were downloaded from the IMG/M database (April 2020). These genomes were recently assembled from public metagenomes as part of a wider effort to generate a genomic catalogue of Earth's microbiome [280]. Finally, 6 metagenome-assembled genomes were obtained as part of the Microbial Dark Matter MDM-II project. CheckM [294] was utilized for estimation of genome completeness, strain heterogeneity, and genome contamination. Only genomes with >70% completion and <10% contamination (n=108) were retained for further analysis (Tables S1, S2). MAGs were classified as high-, or medium-quality drafts based on the criteria set forth by [37]. The utilization of all publicly available genomes through prior individual efforts, as well as the global comprehensive Earth Microbiome collection ensures the global scope of the survey conducted. Continuous addition of new datasets would certainly increase the number of available high-quality Binatota MAGs in the future.

### 3.6.2 Phylogenetic analysis

Taxonomic classifications followed the Genome Taxonomy Database (GTDB) release r89 [292, 296], and were carried out using the classify_workflow in GTDB-Tk [65] (v1.1.0). Phylogenomic analysis utilized the concatenated alignment of a set of 120 single-copy marker genes [292, 293] generated by the GTDB-Tk. Maximum-likelihood phylogenomic tree was constructed in RAxML [369] (with a cultured representative of the phylum Deferrisomatota as the outgroup). SSU rRNA gene-based phylogenetic analysis was also conducted using 16S rRNA gene sequences extracted from genomes using RNAmmer [222]. Putative taxonomic ranks were deduced using average amino acid identity (AAI; calculated using AAI calculator ([http://enve-omics.ce.gatech.edu/]), with the arbitrary cutoffs 56%, and 68% for family, and genus, respectively.

### 3.6.3 Annotation

Protein-coding genes in genomic bins were predicted using Prodigal [179]. For initial prediction of function, pangenomes were constructed for each order in the phylum Binatota separately using PIRATE [22] with percent identity thresholds of [40, 45, 50, 55 ,60, 65, 70, 75, 80, 90], a cd-hit step size of 1, and cd-hit lowest percent id of 90. The longest

sequence for each PIRATE-identified allele was chosen as a representative and assembled into a pangenome. These pangenomes were utilized to gain preliminary insights on the metabolic capacities and structural features of different orders. BlastKOALA [197] was used to assign protein-coding genes in each of the pangenomes constructed to KEGG orthologies (KO), which were subsequently visualized using KEGG mapper [196]. Analysis of specific capabilities and functions of interest was conducted on individual genomic bins by building and scanning hidden markov model (HMM) profiles. All predicted protein-coding genes in individual genomes were searched against custom-built HMM profiles for genes encoding C1, alkanes, and fatty acids metabolism, C1 assimilation, [NiFe] hydrogenases, electron transport chain complexes, and carotenoid and chlorophyll biosynthesis. To build the HMM profiles, Uniprot reference sequences for all genes with an assigned KO number were downloaded, aligned using Clustal-omega [354], and the alignment was used to build an HMM profile using hmmbuild (HMMER 3.1b2). For genes not assigned a KO number (e.g. alternative complex III genes, different classes of cytochrome c family, cytochrome P450 medium-chain alkane hydroxylase cyp153, methanol dehydrogenase MNO/MDO family), a representative protein was compared against the KEGG Genes database using Blastp and significant hits (those with e-values $<$ e-80) were downloaded and used to build HMM profiles as explained above. The custom-built HMM profiles were then used to scan the analyzed genomes for significant hits using hmmscan (HMMER 3.1b2) with the option -T 100 to limit the results to only those profiles with an alignment score of at least 100. Further confirmation was achieved through phylogenetic assessment and tree building procedures, in which potential candidates identified by hmmscan were aligned to the reference sequences used to build the custom HMM profiles using Clustal-omega [354], followed by maximum likelihood phylogenetic tree construction using FastTree [309]. Only candidates clustering with reference sequences were deemed true hits and were assigned to the corresponding KO.

### 3.6.4 Search for photosynthetic reaction center

Identification of genes involved in chlorophyll biosynthesis in Binatota genomes prompted us to search the genomes for photosynthetic reaction center genes. HMM profiles for Reaction Center Type 1 (RC1; PsaAB), and Reaction Center Type 2 (RC2; PufLM and $PsbD_1D_2$) were obtained from the pfam database (pfam00223 and pfam00124, respectively). Additionally, HMM profiles were built for PscABCD (Chlorobia-specific), PshA/B (Heliobacteria-specific) [287], as well as the newly identified Psa-like genes from Chloroflexota [385]. The HMM profiles were used to search Binatota genomes for potential hits using hmmscan. To guard against overlooking a distantly related reaction center, we relaxed our homology criteria (by not including -T or -E options during the hmmscan). An additional search using a structurally-informed reaction center alignment [287, 341] was also performed. The best potential hits were modeled using the SWISS-MODEL homology modeler [396] to check for veracity. Since the core subunits of Type 1 RC proteins are predicted to have 11 transmembrane $\alpha$-helices, while type 2 RC are known to contain five transmembrane helices [5, 167], we also searched for all predicted proteins harboring either 5 or 11 transmembrane domains using TMHMM [215]. All identified 5- or 11-helix-containing protein-coding sequences were searched against GenBank protein nr database to identify and exclude all sequences with a predicted function. All remaining 5- or 11-helix-containing proteins with no predicted func-

tion were then submitted to SWISS-MODEL homology modeler using the automated mode to predict homology models.

### 3.6.5 Classification of [NiFe] hydrogenase sequences

All sequences identified as belonging to the respiratory $O_2$-tolerant $H_2$-uptake [NiFe] hydrogenase large subunit (HyaA) were classified using the HydDB web tool [376].

### 3.6.6 Particulate methane monooxygenase 3D model prediction and visualization

SWISS-MODEL [396] was used to construct pairwise sequence alignments of predicted Binatota particulate methane monooxygenase with templates from ***Methylococcus capsulatus* str. Bath (pdb: 3RGB), and for predicting** tertiary structure models. Predicted models were superimposed on the template **enzyme** in PyMol (Version 2.0 Schrödinger, LLC). Modelling of the active site was conducted similarly. The dicopper-binding site proposed for *Methylococcus capsulatus* str. Bath pMMO [357] (pdb: 3RGB) was used. Alignment of Binatota PmoB sequences with reference *Methylococcus capsulatus* str. Bath PmoB was performed with Clustal-omega [354], and visualized using **the ENDscript server** [330].

### 3.6.7 Ecological distribution of Binatota

We queried 16S rRNA sequence databases using representative 16S rRNA gene sequences from six out of the seven Binatota orders (order UBA12015 genome assembly did not contain a 16S rRNA gene). Two databases were searched: 1. GenBank nucleotide (nt) database (accessed in July 2020) using a minimum identity threshold of 90%, ≥80% subject length alignment for near full-length query sequences or ≥80% query length for non-full-length query sequences, and a minimum alignment length of 100 bp, and 2. The IMG/M 16S rRNA public assembled metagenomes using a cutoff e-value of $1e^{-10}$, percentage similarity ≥ 90%, and either ≥80% subject length for full-length query sequences or ≥80% query length for non-full-length query sequences. Hits satisfying the above criteria were further trimmed after alignment to the reference sequences from each order using Clustal-omega and inserted into maximum likelihood phylogenetic trees in FastTree (v 2.1.10, default settings). The ecological distribution for each of the Binatota orders was then deduced from the environmental sources of their hits. All environmental sources were classified according to the GOLD ecosystem classification scheme [266]. We also queried the datasets from which the Binatota MAGs were binned using the sequence of their ribosomal protein S3. We estimated their relative abundance as the number of reads mapped to contigs with a Binatota ribosomal protein S3 as a percentage to the number of reads mapped to all contigs encoding a ribosomal protein S3 gene. More details on the specifics of the search are in Table S1 footnotes.

## 3.7 Data availability

Genomic bins, predicted proteins, and extended data for Figures 2-3, 5-6, S2-S4, and for Figures 7a-f and S5b-g are available at https://github.com/ChelseaMurphy/Binatota. Maximum likelihood trees (Figure 1 and Figure S5a) can be accessed at: https://itol.embl.de/shared

/1WgxEjrQfEYWk. Maximum likelihood trees for chlorophyll biosynthesis genes are available at https://itol.embl.de/shared/34y3BUHcQd7Lh.

### 3.8 Acknowledgements

This work has been supported by NSF grants 2016423 (to NHY and MSE), 1441717 and 1826734 (to RS). We thank Dr. Kevin Redding (Arizona State University) for helpful discussions. Work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231. J.R.S. is supported by NASA Astrobiology Rock Powered Life and was granted U.S. Forest Service permit #MLD15053 to conduct field work on Cone Pool and the Little Hot Creek, Mammoth Lakes, California. Work on the Paint Pots and Dewar Creek sites was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC DG 2019-06265) and facilitated by Parks Canada, BC Parks, and the Ktunaxa Nation. Thanks to students and participants of the 2014 – 2016 International Geobiology Course for research works on Cone Pool.

This work is published and available in full online (DOI: 10.1128/mBio.00985-21).

### 3.9 Figures & Tables

Supplementary Figures and Tables can be viewed online at:
https://journals.asm.org/doi/10.1128/mBio.00985-21

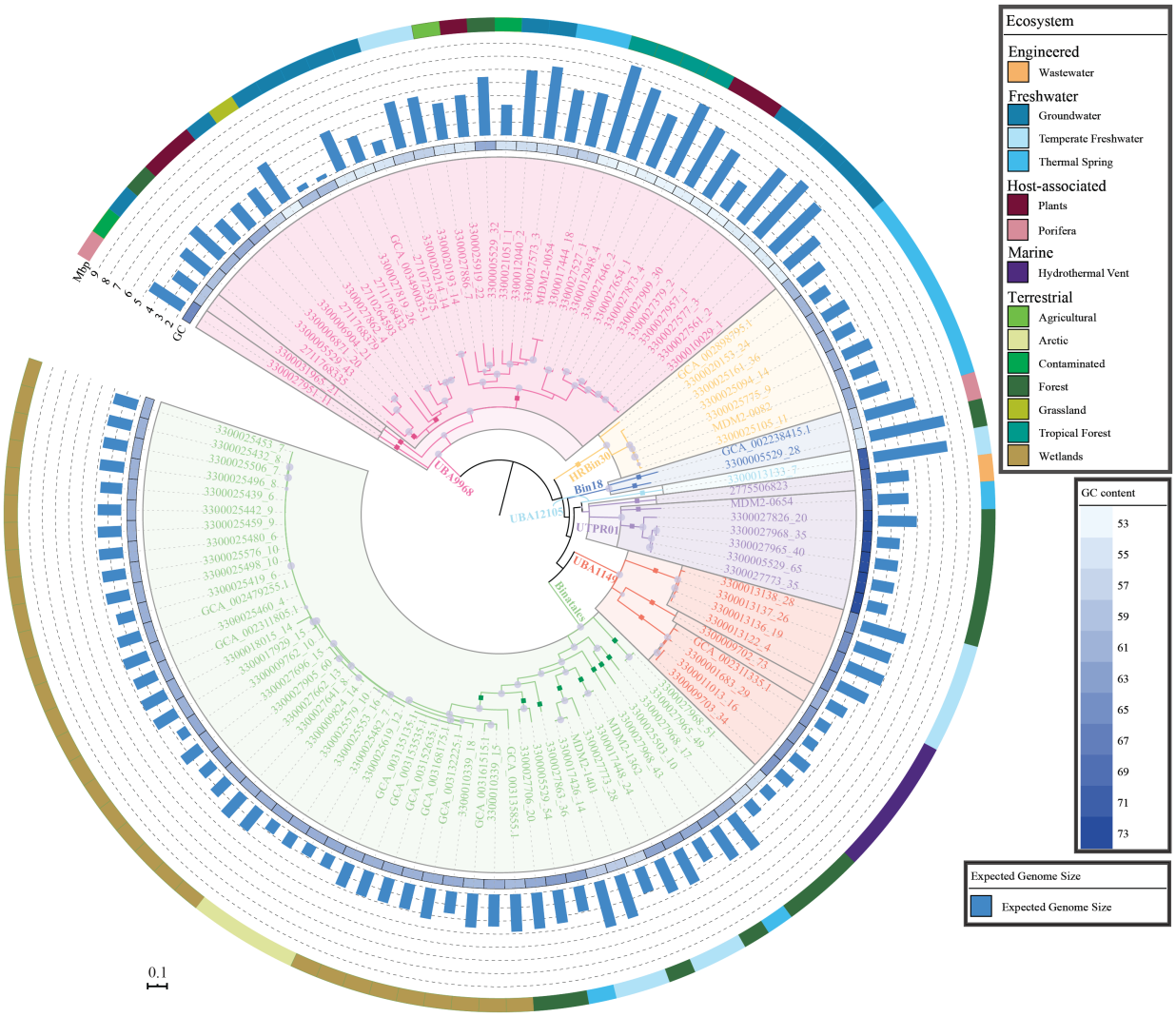



Figure 3.1: Phylogenomic relationship between analyzed Binatota genomes. The maximum-likelihood tree was constructed in RAxML from a concatenated alignment of 120 single-copy marker genes. The tree was rooted using *Deferrisoma camini* (GCA_000526155.1) as the outgroup (not shown). Orders are shown as colored wedges: UBA9968, pink; HRBin30, tan; Bin18, blue; UBA12105, cyan; UTPRO1, purple; UBA1149, orange; and Binatales, green. Within each order, families are delineated by gray borders and genera are shown as colored squares on the branches. Bootstrap values are shown as purple bubbles for nodes with ≥70% support. The tracks around the tree represent (innermost-outermost) G+C content (with a heatmap that ranges from 53% [lightest] to 73% [darkest]), expected genome size (bar chart), and classification of the ecosystem from which the genome originated. All genomes analyzed in this study were >70% complete and <10% contaminated. Completion/contamination percentages and individual genomes assembly size are shown in Tables S2 and S3, respectively.

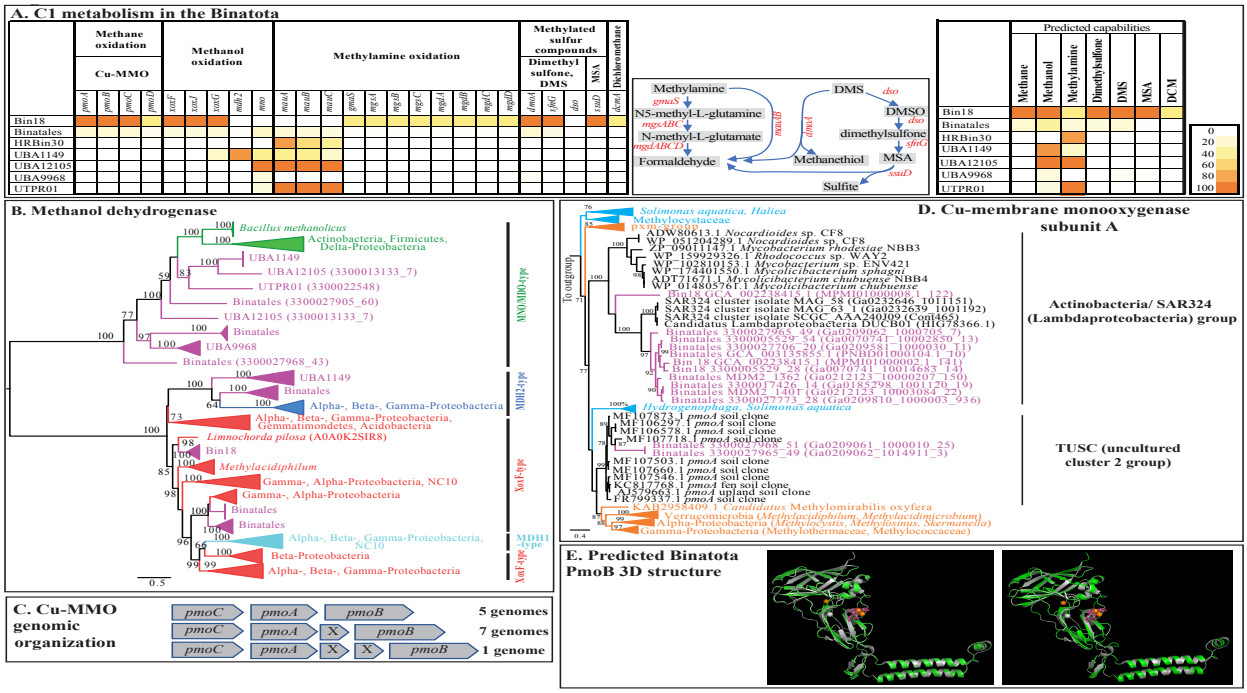



Figure 3.2: $C_1$ substrate degradation capacities in the Binatota. **(A)** Heatmap of the distribution of various $C_1$ oxidation genes in Binatota genomes from different orders. The heatmap colors (as explained in the key) correspond to the percentage of genomes in each order carrying a homologue of the gene in the column header. Pathways involving more than one gene for methylamine and methylated sulfur compounds degradation are shown next to the heatmap. To the right, the per-order predicted $C_1$ oxidation capacity is shown as a heatmap with the colors corresponding to the percentage of genomes in each order where the full degradation pathway was detected for the substrate in the column header. These include *pmoABC* for methane, *xoxFJG*, *mdh2*, and/or *mno* for methanol, *mau* and/or indirect glutamate pathway for methylamine, *sfnG* and *ssuD* for dimethylsulfone, *dso*, *sfnG*, and *ssuD* or *dmoA* for dimethylsulfide (DMS), *ssuD* for methane sulfonic acid (MSA), and *dcmA* for dichloromethane (DCM). CuMMO, copper membrane monooxygenase with subunits A, B, C, and D; XoxF-type (*xoxF*, *xoxJ*, *xoxG*), MDH2-type (*mdh2*), and MNO/MDO-type (*mno*) methanol dehydrogenases; direct oxidation methylamine dehydrogenase (*mauABC*); indirect glutamate pathway (*gmaS*, $\gamma$-glutamylmethylamide synthase; *mgsABC*, N-methyl-L-glutamate synthase; *mgdABCD*, methylglutamate dehydrogenase); dimethylsulfide (DMS) monooxygenase (*dmoA*); dimethyl sulfone monooxygenase (*sfnG*); dimethylsulfide monooxygenase (*dso*); alkane sulfonic acid monooxygenase (*ssuD*); and dichloromethane dehalogenase (*dcmA*). DMSO, dimethyl sulfoxide. **(B)** Maximum-likelihood phylogenetic tree highlighting the relationship between Binatota methanol dehydrogenases in relation to other methylotrophic taxa. Bootstrap support (from 100 bootstraps) is shown for branches with >50% bootstrap support. **(C)** Organization of CuMMO genes in Binatota genomes and the number of genomes where each organization was observed. X, hypothetical protein. **(D)** Maximum-likelihood tree highlighting the relationship between Binatota *pmoA* genes to methanotrophic taxa and environmental amplicons. Bootstrap support (100 bootstraps) is shown for branches with >50% bootstrap support. Sequences from Binatota genomes (shown as order followed by bin name and then PmoA protein ID in parentheses) are in magenta and fall into two clusters: *Actinobacteria*/SAR324 cluster and TUSC uncultured cluster 2. Clusters from previously studied CuMMOs known to reduce methane are in orange, while those known to reduce short-chain alkanes but not methane are in cyan. The tree was rooted using the amoA sequence of "*Candidatus* Nitrosarchaeum limnium SFB1" (EGG41084.1) as an outgroup. **(E)** Predicted PmoB 3D structure (gray) from a cluster 2 TUSC-affiliated Binatota genome (genome 3300027968_51, left), and an Actinobacteria/SAR324-affiliated Binatota genome (genome GCA_002238415.1, right) both superimposed on PmoB from the model methanotroph *Methylococcus capsulatus* strain Bath (PDB: 3RGB) (green) with global model quality estimation (GMQE) scores of 0.73 and 0.62, respectively.

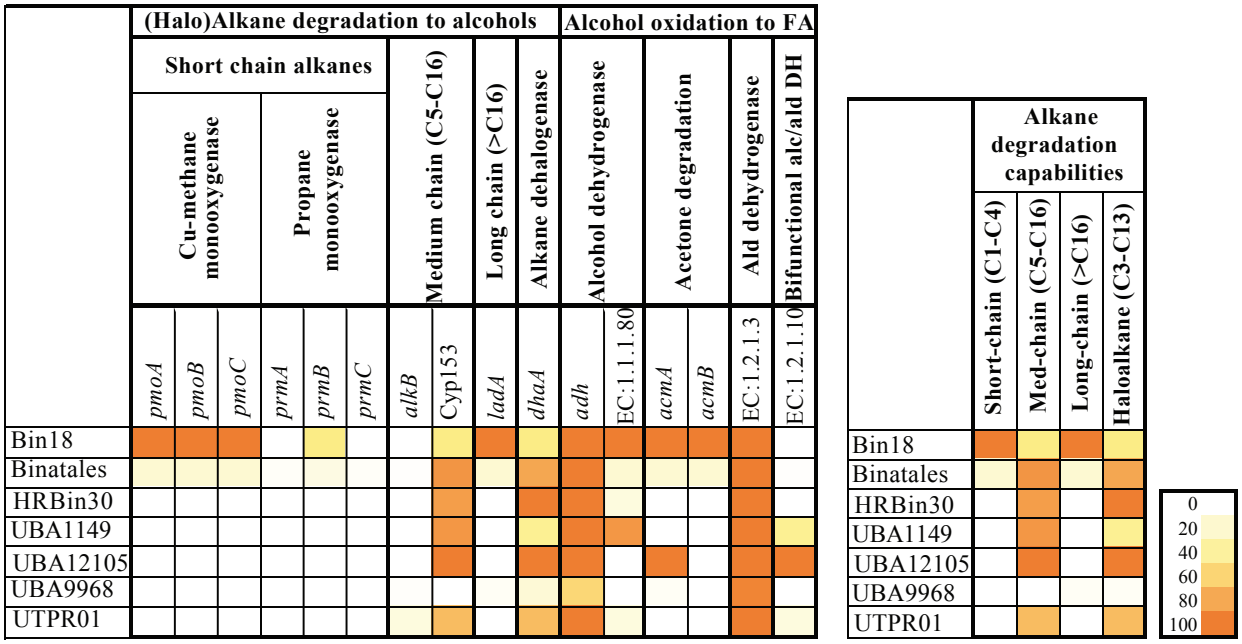



Figure 3.3: Heatmap of the distribution of (halo)alkane degradation to alcohol in Binatota genomes. The heatmap colors (as explained in the key) correspond to the percentage of genomes in each order carrying a homologue of the gene in the column header. The per-order predicted alkane-degradation capacity is shown to the right as a heatmap with the colors corresponding to the percentage of genomes in each order where the full degradation pathway was detected for the substrate in the column header. These include CuMMO and/or *prmABC* for short-chain alkanes, *alkB* or Cyp153 for medium-chain alkanes, *ladA* for long-chain alkanes, and *dhaA* for haloalkanes. Ald, aldehyde.

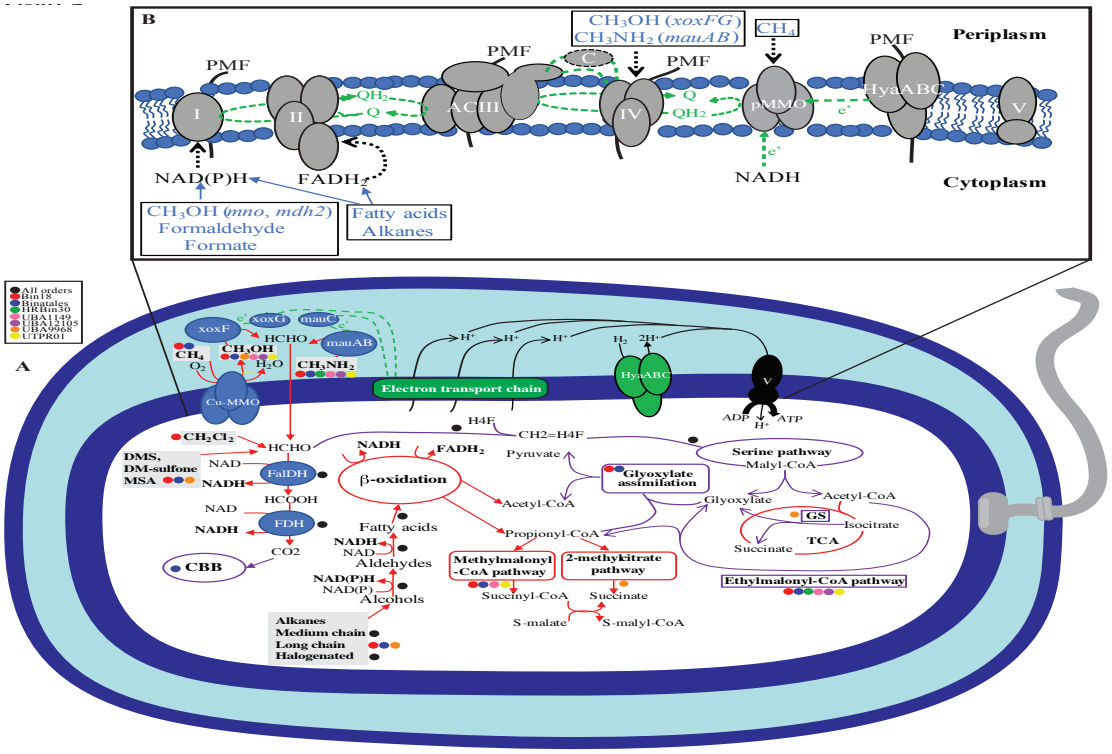



Figure 3.4: **(A)** Cartoon depicting different metabolic capabilities encoded in the Binatota genomes with capabilities predicted for different orders shown as colored circles as shown in the legend. Enzymes for $C_1$ metabolism are shown in blue and include the copper membrane monooxygenases (CuMMOs), methanol dehydrogenase (*xoxFG*), and methylamine dehydrogenase (*mauABC*), as well as the cytoplasmic formaldehyde dehydrogenase (FalDH) and formate dehydrogenase (FDH). Electron transport chain is shown as a green rectangle. Electron transfer from periplasmic enzymes to the ETC is shown as dotted green lines. The sites of proton extrusion to the periplasm are shown as black arrows, as is the F-type ATP synthase. Carbon dissimilation routes are shown as red arrows, while assimilatory routes are shown as purple arrows. Details of the assimilatory pathways are shown in Fig. S2 and S3. Reducing equivalents potentially fueling the ETC [NAD(P)H and $FADH_2$] are shown in boldface. All substrates predicted to support growth are shown in boldface within gray boxes. A flagellum is depicted, the biosynthetic genes of which were identified in genomes belonging to all orders except Bin18, HRBin30, and UBA1149. The cell is depicted as rod-shaped based on the identification of the rod shape determining gene *rodA* in all genomes and the rod-shape determining genes *mreB* and *mreC* in genomes from all orders except UBA1149. The inset on top **(B)** details the electron transport chain in the Binatota with all electron transfer complexes (I, II, ACIII, IV) embedded in the inner membrane, along with the particulate methane monooxygenase (pMMO) and the $H_2$-uptake [NiFe] hydrogenase (HyaABC). All genomes also encoded an F-type ATP synthase complex (V). Substrates potentially supporting growth are shown in blue with predicted entry points to the ETC shown as dotted black arrows. Sites of proton extrusion to the periplasm and proton motive force (PMF) creation are shown as solid black lines, while sites of electron (e') transfer are shown as dotted green lines. Three possible physiological reductants are shown for pMMO (as dotted green arrows): the quinone pool coupled to ACIII, NADH, and/or some of the reduced quinones generated through $H_2$ oxidation by HyaABC. Abbreviations: CBB, Calvin Benson Bassham cycle; FalDH, NAD-linked glutathione-independent formaldehyde dehydrogenase, *fdhA*; FDH, NAD-dependent formate dehydrogenase (EC: 1.17.1.9); Fum, fumarate; GS, glyoxylate shunt; $H_4F$, tetrahydrofolate; HyaABC, type I respiratory $O_2$-tolerant $H_2$-uptake [NiFe] hydrogenase; *mauABC*, methylamine dehydrogenase; CuMMO. copper membrane monooxygenases; *xoxFG*, xoxF-type methanol dehydrogenase; succ, succinate; TCA, tricarboxylic acid cycle; V, F-type ATP synthase (EC: 7.1.2.2 7.2.2.1).

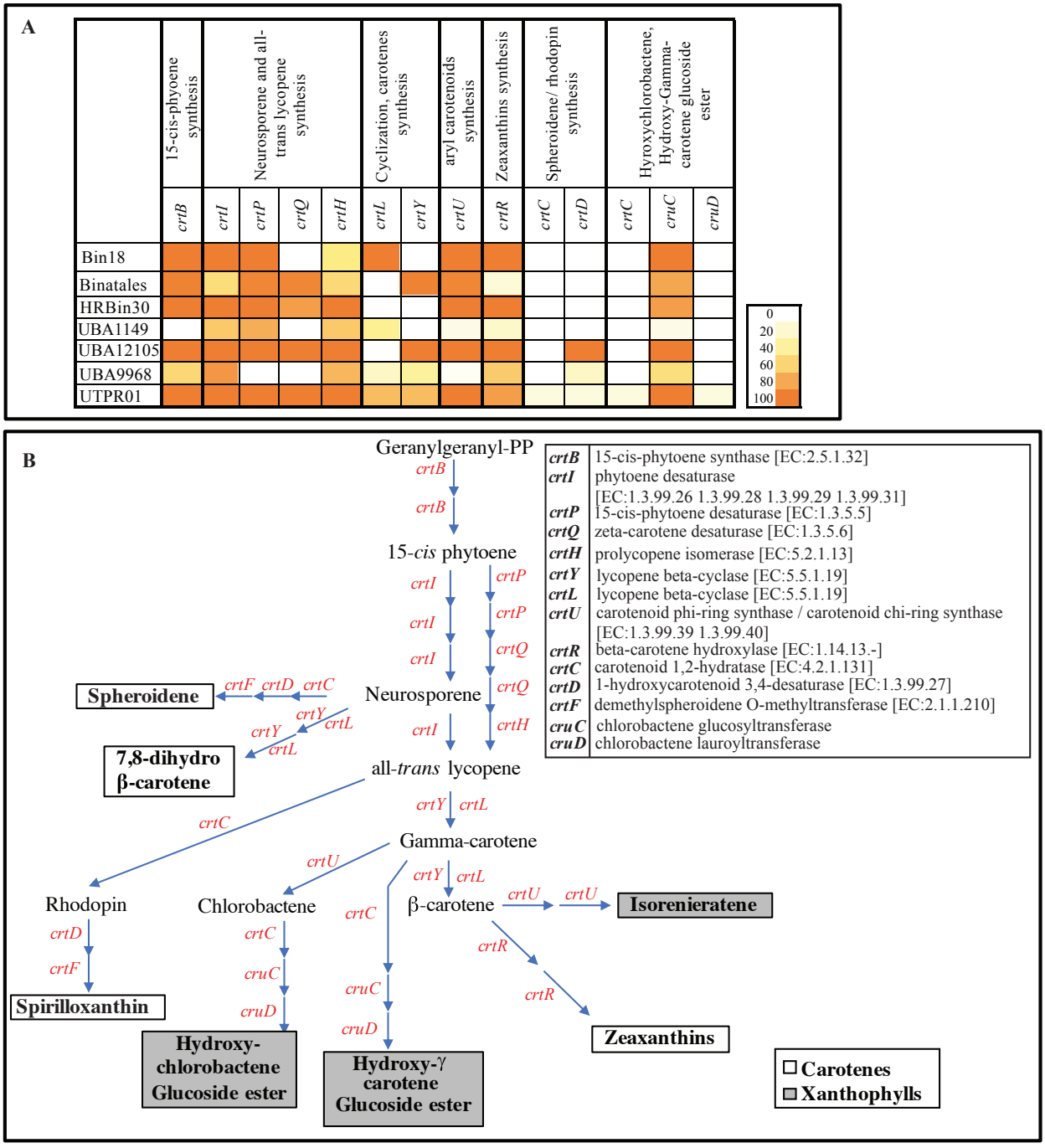



Figure 3.5: Carotenoids biosynthesis capabilities in Binatota genomes. **(A)** Distribution of carotenoid biosynthesis genes in the Binatota genomes. The heatmap colors (as explained in the key) correspond to the percentage of genomes in each order encoding a homologue of the gene in the column header. **(B)** Carotenoid biosynthesis scheme in Binatota based on the identified genes. Genes encoding enzymes catalyzing each step are shown in red and their descriptions with EC numbers are shown to the right. Binatota genomes encode the capability to biosynthesize both exclusively hydrocarbon carotenes (white boxes) or the oxygenated xanthophylls (gray boxes).

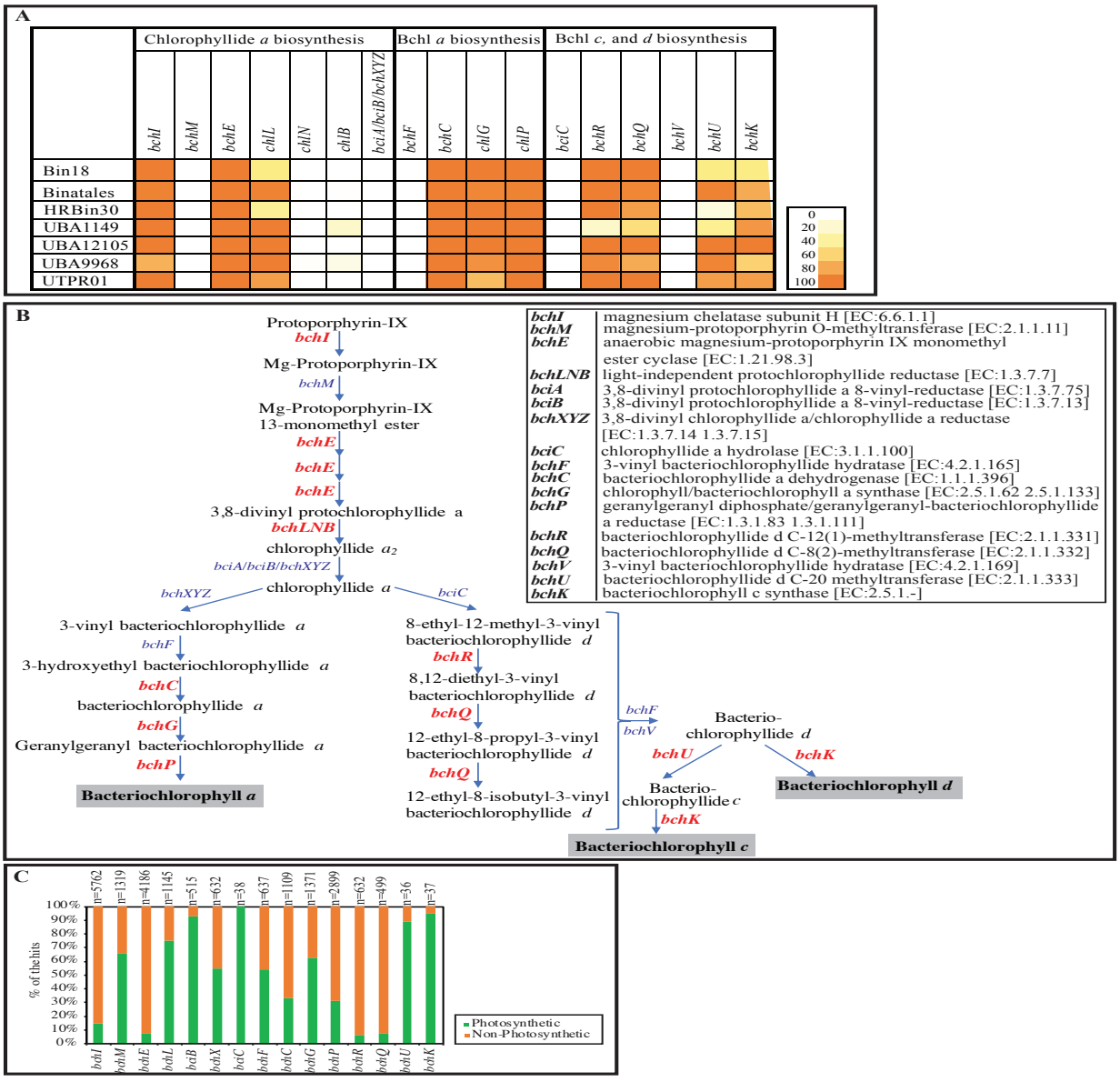



Figure 3.6: Bacteriochlorophylls biosynthesis genes encountered in Binatota genomes studied suggesting an incomplete pathway for bacteriochlorophyll *a*, *c*, and/or *d* biosynthesis. **(A)** Distribution of chlorophyll biosynthesis genes in Binatota genomes. The heatmap colors (as explained in the key) correspond to the percentage of genomes in each order carrying a homologue of the gene in the column header. **(B)** Bacteriochlorophylls biosynthesis pathway. Genes identified in at least one Binatota genome are shown in red boldface text, while these with no homologues in the Binatota genomes are shown in blue text. Gene descriptions with EC numbers are shown to the right of the figure. **(C)** Distribution patterns of bacteriochlorophyll biosynthesis genes. The search was conducted in the functionally annotated bacterial tree of life AnnoTree [253] using single KEGG orthologies implicated in chlorophyll biosynthesis. Gene names are shown on the x axis, total number of hits is shown above the bar for each gene, and the percentage of hits in genomes from photosynthetic (green) versus nonphotosynthetic (orange) genera are in the stacked bars.

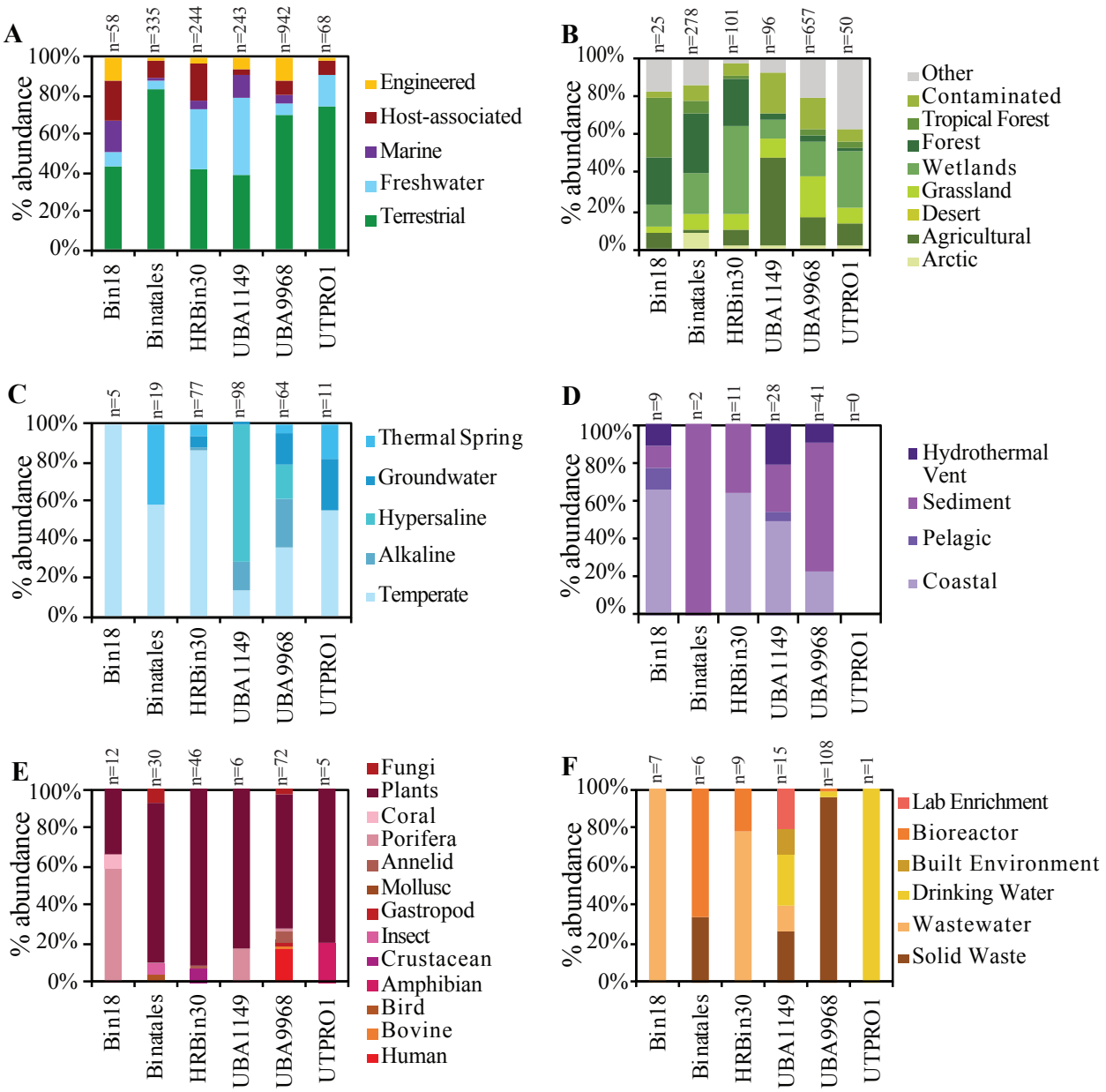



Figure 3.7: Ecological distribution of Binatota-affiliated 16S rRNA sequences in GenBank nt database. Binatota orders are shown on the $x$ axis, while percent abundance in different environments (classified based on the GOLD ecosystem classification scheme) is shown on the $y$ axis **(A)**. Further subclassifications for each environment are shown for **(B)** terrestrial, **(C)** freshwater, **(D)** marine, **(E)** host-associated, and (F) engineered environments. The total number of hit sequences for each order is shown above the bar graphs. Details, including GenBank accession number of hit sequences, are shown in Extended Data 2. Order UBA12105 genome assembly did not contain a 16S rRNA gene, so this order is not included in the analysis.

# CHAPTER IV

# GENOMIC CHARACTERIZATION OF THREE NOVEL DESULFOBACTEROTA CLASSES EXPAND THE METABOLIC AND PHYLOGENETIC DIVERSITY OF THE PHYLUM

## 4.1 Originality-Significance Statement

Culture-independent diversity surveys conducted in the last three decades have clearly demonstrated that the scope of microbial diversity is much broader than that inferred from isolation efforts. Multiple reasons have been put forth to explain the refractiveness of a wide range of the earth's microbiome to isolation efforts. Documenting the scope of high-rank phylogenetic diversity on earth, as well as deciphering and documenting the metabolic capacities, physiological preferences, and putative ecological roles of these yet-uncultured lineages represents one of the central goals in current microbial ecology research. Recent efforts to assemble genomes from metagenomes have provided invaluable insights into these yet-uncultured lineages. This study expands our knowledge of the phylum Desulfobacterota through the characterization of 30 genomes belonging to three novel classes. The analyzed genomes were either recovered from Zodletone Spring in southwestern Oklahoma in this study, or recently binned from public metagenomes as part of the Global Earth Microbiome initiative. Our results expand the high-rank diversity within the bacterial tree of life by describing three novel classes within the phylum Desulfobacterota, document the utilization of multiple metabolic processes (iron oxidation, aromatic hydrocarbon degradation, reduction of sulfur-cycling intermediates), and features (coenzyme M biosynthesis, and pigmentation) as salient characteristics in these novel Desulfobacterota classes.

## 4.2 Summary

We report on the genomic characterization of three novel classes in the phylum Desulfobacterota. One class (proposed name *Candidatus* "Anaeroferrophillalia") was characterized by heterotrophic growth capacity, either fermentatively or utilizing polysulfide, tetrathionate or thiosulfate as electron acceptors. In the absence of organic carbon sources, autotrophic growth via the Wood Ljungdahl (WL) pathway and using hydrogen or Fe(II) as an electron donor is also inferred for members of the "Anaeroferrophillalia". The second class (proposed name *Candidatus* "Anaeropigmentia") was characterized by its capacity for growth at low oxygen concentration, and the capacity to synthesize the methyl/alkyl carrier CoM, an ability that is prevalent in the archaeal but rare in the bacterial domain. Pigmentation is inferred from the capacity for carotenoid (lycopene) production. The third class (proposed name *Candidatus* "Zymogenia") was characterized by fermentative heterotrophic growth ca-

pacity, broad substrate range, and the adaptation of some of its members to hypersaline habitats. Analysis of the distribution pattern of all three classes showed their occurrence as rare community members in multiple habitats, with preferences for anaerobic terrestrial, freshwater, and marine environments over oxygenated (e.g. pelagic ocean and agricultural land) settings. Special preference for some members of the class *Candidatus* "Zymogenia" for hypersaline environments such as hypersaline microbial mats and lagoons was observed.

## 4.3 Introduction

Traditional approaches to characterize bacterial and archaeal taxa have long hinged upon isolation procedures. Despite decades of such efforts, much of the microbial world remains uncultured (Hug et al., 2016). Culture-independent surveys utilizing the 16S rRNA gene as a phylogenetic marker have long been utilized to characterize yet-uncultured microbial diversity. However, while useful for documenting the identity, relative abundance, and distribution patterns of microorganisms, such surveys provide limited information on community interactions, metabolic inferences, or microbial growth rates [28, 44, 327]. The development and wide-scale utilization of genome-resolved metagenomics and single cell genomic approaches have yielded a wealth of metagenome-assembled genome (MAGs) and single-cell amplified genome (SAGs) assemblies [327, 404]. In addition to successfully recovering representative genomes of uncultured lineages previously defined by 16S rRNA gene surveys [404], such efforts have also expanded the bacterial tree of life by recovering representatives of novel lineages that previously eluded detection even by culture-independent diversity surveys [10, 173, 229, 327].

Further, the accumulation of MAGs and SAGs has been leveraged for enacting phylogenomic-based taxonomic schemes that encompass both cultured and uncultured microorganisms [291]. The genome taxonomy database (GTDB) utilizes 120 bacterial single-copy genes, average nucleotide identity (ANI), alignment fractionation, and specific quality controls to provide a detailed (phylum to species) genome-based taxonomy [65]. The current release (r95) encompasses 111 phyla, 327 classes, 917 orders, 2,282 families, 8,778 genera, and 30,238 species. While broadly agreeing with organismal and 16S rRNA gene-based outlines, e.g. [315], the GTDB proposes several name and rank changes. For example, the Gram-positive Firmicutes are proposed to constitute seven different phyla (Firmicutes A-G). Within the Proteobacteria, the class Betaproteobacteria was reclassified as an order within the class Gammaproteobacteria, and the Epsilonproteobacteria was elevated to a new phylum (Campylobacterota) [391].[291]

Perhaps nowhere was the effect of genomics-based taxonomy more profound than in the class Deltaproteobacteria. The Deltaproteobacteria encompassed anaerobic respiratory and fermentative/syntrophic lineages, as well as the Myxobacteria, *Bdellovibrio*-like predators, and aquatic oligotrophs. The disparate physiological preferences, metabolic capacities, and lifestyles within the Deltaproteobacteria have long been noted, and prior efforts based on single and concatenated gene phylogenies have reported its polyphyletic nature [147]. The recent genome-based taxonomy in GTDB r95 has proposed splitting this group into 16 phyla, including 4 distinct cultured phyla: Desulfobacterota, Myxococcota, Bdellovibrionota, and SAR324 [390].

The recently enacted phylum Desulfobacterota encompasses sulfate-reducing and related

fermentative and syntrophic lineages [184, 240] previously constituting the bulk of strict anaerobes within previously classified Deltaproteobacteria. GTDB r95 lists 20 classes, 31 orders, 119 families, and 279 genera, of which 12, 14, 38, and 86 contain cultured representatives, respectively. Cultured members of the Desulfobacterota have been identified in a plethora of marine, freshwater, terrestrial, and engineered ecosystems exhibiting wide ranges of salinities, pH, and temperatures [270]. Cultured Desulfobacterota show preference for anoxic conditions, and many utilize sulfate, sulfite, thiosulfate, elemental sulfur, or iron (or combinations thereof) as the terminal electron acceptor in respiratory and/or disproportionation processes [142, 219, 242, 355].

During a broad effort to characterize the yet-uncultured diversity within Zodletone spring, a sulfide and sulfur-rich spring in southwestern Oklahoma, using genome-resolved metagenomics, we recovered multiple MAGs that appear to represent distinct novel classes within the phylum Desulfobacterota. In addition, a recent global effort for binning genomes from publicly available metagenomic datasets [280] yielded additional MAGs belonging to these novel Desulfobacterota classes. Here, we report on the genomic characteristics; inferred metabolic capacities, physiological preferences, structural features, and ecological distribution of three novel additional classes within the Desulfobacterota.

## 4.4 Results

### 4.4.1 Three novel classes within the Desulfobacterota

Thirty Desulfobacterota MAGs binned from Zodletone spring sediments and 12 other locations (Table 1, Table S1) clustered into three distinct clades comprising 7, 17, and 6 genomes, that were unaffiliated with any of the 20 currently recognized Desulfobacterota classes in GTDB r95 taxonomic outline (Table 1, Table S1, Figure 1). Average amino acid identity (AAI) and shared gene content (SGC) values between these genomes and representative genomes from all other classes within the Desulfobacterota ranged between 41.63% ± 0.96% – 42.71% ± 1.04% (AAI), and 44.52% ± 4.28% – 51.61% ± 4.37% (SGC), confirming their distinct suggested position as novel classes within the Desulfobacterota phylum (Table S1). Further, the obtained Relative Evolutionary Divergence (RED) values of 0.38 – 0.42 confirmed the distinct class-level designation for all three lineages (Table S1). 16S rRNA gene sequences extracted from representative genomes placed all three groups as members of the "uncultured Deltaproteobacteria" bin within the class Deltaproteobacteria, Phylum Proteobacteria using the RDP-II taxonomic outline (Table S1). In SILVA taxonomic outline release 138.1 [315], these clades were classified as unclassified members of the order Desulfobacterales (clade 1), members of the phylum Sva085 (clade 2), and as uncultured members of the Desulfobacterota phylum (clade 3) (Table S1). We hence propose accommodating these 30 genomes into three distinct classes, for which the following names are proposed based on defining metabolic characteristics predicted from their genomes as described below: *Candidatus* "Anaeroferrophillalia" (order Anaeroferrophillales, family Anaeroferrophillacea), with the MAG assembly Zgenome_940 (*Anaeroferrophillus wilburensis*) serving as the type material (GenBank assembly accession number JAFGSY000000000) (the name reflects its preference for anaerobic environments and predicted capacity to utilize Fe(II) as a supplementary electron donor in absence of organic substrates); *Candidatus* "Anaeropigmentia" (order

Anaeropigmentiales, family Anaeropigmentiaceae), with the MAG assembly 3300022855_4 (*Anaeropigmentus antarcticus*) serving as the type material (IMG assembly accession number 3300022855_4) (the name reflects its preference for anaerobic environments and predicted capacity for pigment biosynthesis); and *Candidatus* "Zymogenia" (order Zymogeniales, family Zymogeniaceaea), with the MAG assembly Zgenome_24 (*Zymogenus saltonus*) serving as the type material (GenBank assembly accession number JAFGIX000000000) (the name reflects its preference for a fermentative mode of metabolism (zymo: Greek for digestion and fermentation). The representative type material MAG was chosen based on the MAG quality deduced from the % completeness (>90%), % contamination (<5%), and the presence of a rRNA operon in the assembly (Table S2) [37]. These three classes were further classified into three orders, 6 families, and 10 genera based on the intra-class AAI values (Table S1). Below, we provide a more detailed analysis of the inferred metabolic capacities, physiological preferences, and ecological distribution of each of these three classes.

### 4.4.2 Structural, physiological, and metabolic features

**Class "Anaeroferrophillalia"**

**General genomic features:** Genomes belonging to Class "Anaeroferrophillalia" possess average sized genomes (2.80 ± 0.33 Mbp), GC content (52.66% ± 5.63%), and gene length (937.07 ± 37.07 bp) (Table 1). Structurally, members are predicted to have a Gram-negative cell wall based on the possession of lipopolysaccharide (LPS) biosynthesis-encoding genes, and lack of genes encoding the pentaglycine linkage of peptidoglycan. The presence of the rod-shape determining genes *rodA*/*mreB* and genes encoding flagellar assembly suggest rod-shaped flagellated cells. Defense mechanisms include CRISPR defense systems and Type I restriction endonucleases (Table S3). No evidence for special intracellular structures, e.g. bacterial microcompartments, nanocompartments, or magnetosomes, was identified (Table S3).

**Physiological features:** Members of class "Anaeroferrophillalia" appear to be strict anaerobes, based on the absence of respiratory cytochrome C oxidase (complex IV) components, the presence of the oxygen-limited cytochrome *bd* complex, and the identification of the oxidative stress enzymes catalase, rubrerythrin, rubredoxin, alkylhydroperoxide reductase, and peroxidase (Table S3). Osmoadaptive capabilities are predicted via the identification of glycine betaine/proline ABC transporter ProXWV.

**Heterotrophic fermentative capacities:** Genomes of Class "Anaeroferrophillalia" possess robust biosynthetic capacities with few amino acids or cofactors auxotrophies (Table S3). The presence of genes encoding the Embden–Meyerhof–Parnas (EMP) and the non-oxidative pentose phosphate pathway pathway (PPP) indicate heterotrophic growth capabilities. However, a limited number of sugars (glucose, fructose, mannose) appear to support growth (Figure 2A). As well, the capacity to degrade amino acids appears to be limited (Figure 2A, Table S3), and the lack of genes encoding the beta-oxidation pathway precludes potential growth on medium- and long-chain fatty acids. On the other hand, all genomes encode the lactate utilization enzyme D-lactate dehydrogenase (cytochrome) [EC:1.1.2.4],

suggesting the capability to grow on D-lactate (Figure 2A). As well, the pathway for anaerobic benzoate metabolism appears to be present in all genomes, suggesting a specialty in degradation of aromatic compounds (Figure 2A). Pyruvate generated from D-lactate or sugar metabolism could be metabolized to acetyl-CoA via pyruvate ferredoxin oxidoreductase encoded in all genomes, followed by conversion of acetyl-CoA to acetate with concomitant substrate level phosphorylation via the acetate CoA ligase (Figure 2A, Table S3).

**Respiratory capacities:** In addition to fermentative capacities, possible respiratory activities were identified in class "Anaeroferrophillalia". Possible electron donors identified based on genomic analysis include D-lactate via the D-lactate dehydrogenase [EC: 1.1.2.5]. This enzyme has been studied in several sulfate reducers and its physiological electron acceptor was found to be ferricytochrome c3, which could serve as an entry point to an ETS, with the electrons possibly moving to the genomically-encoded Qrc membrane complex (menaquinone reductase), on to the quinone pool and eventually to the terminal electron acceptor. Several hydrogenase-encoding genes were identified in the genomes of class "Anaeroferrophillalia". These include the periplasmic [Ni Fe] HyaABC (HydDB group 1d), predicted to be involved in hydrogenotrophic respiration (as well as other hydrogenases that are predicted to be involved in recycling reduced equivalents as explained below). Hydrogenotrophic respiration would proceed through the periplasmic hydrogenase moving electrons from $H_2$ onto a periplasmic cytochrome C, the Qrc membrane complex, the quinone pool, and eventually to the terminal electron acceptor.

Further, analysis of iron metabolism genes in members of class "Anaeroferrophillalia" genomes indicated their possession of operonic DFE_0448-0451 and DFE_0461-0465 genes, similar to the systems first identified in *Desulfovibrio ferrophilus* (Figure 2B) [96]. In the proposed *D. ferrophilus* model, electrons move from an external source, e.g. insoluble minerals like iron, to an outer membrane cytochrome (encoded by DFE_0450 and DFE_0464, respectively) through a complex of additional heme-containing periplasmic membrane-bound (DFE_0449 and DFE_0461), periplasmic soluble (DFE_0448 and DFE_0462, DFE_0465), or complex stabilizing (DFE_0451 and DFE_0463) cytochromes. Electrons could potentially then pass on to menaquinone [96] and eventually to the terminal electron acceptor. This later ETS is expected to operate possibly under substrate-limiting conditions (for example in absence of D-lactate) as shown before for *D. ferrophilus* [96].

Possible electron acceptors identified include the sulfur cycle intermediates tetrathionate, based on the identification of genes encoding the octaheme tetrathionate reductase (Otr) [265], as well as the guanylyl molybdenum cofactor-containing tetrathionate reductase (TtrABC) [163]. The produced thiosulfate from the action of either Otr or TtrABC could be metabolized through disproportionation, based on the identification of thiosulfate reductase *phsABC* genes (Figure 2B). Polysulfide reduction capability is also predicted based on the identification of genes encoding the membrane-bound molybdoenzyme complex PsrABC [99]. No marker genes suggesting the ability to respire sulfate or sulfite were identified. Nitrate reduction genes were similarly lacking.

**ATP production and recycling reduced equivalents:** The genomes encoded complex I components, NADH dehydrogenase, as well as an $F_1F_o$-ATPase. With an incomplete

oxidative phosphorylation pathway, we predict that the NADH dehydrogenase is possibly coupled to the quinone pool and cytochromes for generation of a proton motive force across the inner membrane that can then be used for ATP synthesis via the $F_1F_o$-ATPase, similar to the model predicted in the sulfate reducer *Desulfovibrio vulgaris* [288]. Alternatively, or concomitantly, a proton-motive force could possibly be generated during the operation of Wood Ljungdahl (WL) pathway, encoded by all Class "Anaeroferrophillalia" genomes, in the homoacetogenic direction. In that case, a membrane-bound mechanism that achieves redox balance between heterotrophic substrate oxidation and the WL function as the electron sink [345, 415] is needed. Candidates for this membrane-bound electron bifurcation mechanism are the membrane-bound [Ni Fe] hydrogenase (Mbh) (HydDB group 4d), which couples reduced ferredoxin (produced via the action of pyruvate ferredoxin oxidoreductase) oxidation to the reduction of protons to $H_2$, with the concomitant export of protons to the periplasm [345, 415]. Recycling of electron carriers would further be achieved by the cytoplasmic [NiFe]-hydrogenase (MvhAGD) plus the heterodisulfide reductase HdrABC, both of which are encoded in the genomes (HydDB group [Ni Fe] 3c) (Figure 2A).

**Autotrophic capacities:** In conditions where inorganic compounds (e.g. ferrous ions or $H_2$) serve as electron donors, we expect autotrophic capacities to be fulfilled via the WL pathway. In that case, the proton gradient-driven phosphorylation (through the ATPase complex) will be the only means for ATP production [275] as no net ATP gain by substrate level phosphorylation (SLP) is achieved via the WL pathway [415]. All genomes encode mechanisms for acetyl-CoA (produced from WL pathway) conversion to pyruvate (pyruvate:ferrodoxin oxidoreductase), reversal of pyruvate kinase (including pyruvate-orthophosphate dikinase [EC: 2.7.9.1], pyruvate-water dikinase [EC: 2.7.9.2], as well as pyruvate carboxylase and PEP carboxykinase (ATP) [EC:4.1.1.49]), and the bifunctional fructose-1,6-bisphosphate aldolase/phosphatase to reverse phosphofructokinase.

### Class "Anaeropigmentia"

**General genomic features:** Members of the Class "Anaeropigmentia" possess relatively large genomes (3.96 ± 0.74 Mbp), with average GC content (47.29% ± 4.55%), and gene length (912.46 ±44.91 bp) (Table 1). Cells are predicted to be Gram-negative rods, with CRISPR systems and type I, and type III restriction endonucleases (Table S3).

**Physiological features:** Class "Anaeropigmentia" genomes encode two distinct pathways for the biosynthesis of the compatible solute trehalose (both from ADP-glucose and glucose via trehalose synthase, as well as from UDP-glucose and glucose via the action of trehalose 6-phosphate synthase and trehalose 6-phosphate phosphatase), and its degradation via alpha, alpha-trehalose phosphorylase. The genomes also encoded the capability for biosynthesis and degradation of the storage molecule starch.

**Heterotrophic capacities:** A heterotrophic lifestyle is predicted, based on the presence of a complete EMP and Entner–Doudoroff (ED) pathways, a complete TCA cycle, and both the oxidative and non-oxidative branches of the PPP. Substrates predicted to support growth include sugars (glucose, fructose, mannose, sorbitol), amino acids (alanine, aspartate,

asparagine, glutamine, glutamate, cysteine, serine), and fatty acids via the beta-oxidation pathway (Table S3, Figure 3A).

**Fermentative capacities:** The capability of class "Anaeropigmentia" to ferment pyruvate is inferred by the presence of various pathways for end product (butanediol, acetate, ethanol, and acetoin) generation (Table S3, Figure 3A).

**Respiratory capacities:** A complete electron transport chain with complexes I (NADH-quinone oxidoreductase), II (succinate dehydrogenase), alternate complex III (encoded by *actABCDEFG)*, and complex IV (cytochrome c oxidase cbb3-type, as well as cytochrome *bd* ubiquinol oxidase), in addition to a V/A-type as well as F-type $H^+/Na^+$-transporting ATPase, were identified suggesting possible utilization of trace amounts of $O_2$ as a terminal electron acceptor by members of Class "Anaeropigmentia". All Class "Anaeropigmentia" genomes encoded a complete WL pathway. Additional ATP production via oxidative phosphorylation following the generation of a proton-motive force during the operation of WL pathway is therefore also predicted. In that case, the *Rhodobacter* nitrogen fixation (RNF) complex encoded in the majority of genomes would re-oxidize reduced ferredoxin at the expense of NAD, with the concomitant export of protons to the periplasm, thus achieving redox balance between heterotrophic substrate oxidation and the WL function as the electron sink [345, 415]. Recycling of electron carriers would further be achieved by the cytoplasmic electron bifurcating mechanism (HydABC) plus MvhAGD-HdrABC, both of which are encoded in the genomes (HydDB groups [Fe Fe] A3, and [Ni Fe] 3c).

**Specialized cofactor biosynthesis:** Interestingly, the complete pathway encoding the phosphoenol pyruvate-dependent coenzyme M (CoM) biosynthesis was identified in all genomes of Class "Anaeropigmentia" (Figure 3A, Table S3). CoM is a hallmark of methanogenic Archaea, where it acts as a terminal methyl carrier that releases methane upon regeneration of its unmethylated state during methanogenesis [377]. However, the utility of CoM in the bacterial domain is less understood. CoM was shown in the bacterial genera *Xanthobacter, Rhodococcus,* and *Mycobacterium* to be involved in propylene degradation as a carrier for a C3 carbon intermediate [6, 78, 218]. Recently, the bacterial CoM biosynthetic cluster was identified in *X. autotrophicus* Py2 [299]. The genes *xcbB1*, *C1*, *D1*, and *E1 encode the bacterial CoM biosynthetic operon, with* only *xcbB1* showing homology to the archaeal CoM biosynthesis gene *comA* [299]. The remainder of the genes *xcbC1*, *D1*, and *E1 are distinct from the archaeal genes comBCDE, and the bacterial biosynthetic pathway proceeds via a different route* [140, 299]. CoM biosynthesis genes identified in Class "Anaeropigmentia" genomes are distinct from the bacterial CoM biosynthesis genes *xcbC1*, *D1*, and *E1* and are indeed archaeal-like. Searching the functionally annotated bacterial tree of life AnnoTree [253] using the combination of KEGG orthologies corresponding to the archaeal CoM biosynthetic cluster *comABCDE* identified their collective presence in only 14 bacterial genomes from the phyla Acidobacteriota, Actinobacteriota, Bacteroidota, Chloroflexota, Desulfobacterota, Desulfobacterota_B, Latescibacterota, and Proteobacteria. Unfortunately, genes encoding additional enzymes required for propylene degradation (alkene monooxygenase, 2-hydroxypropyl-CoM lyase, 2-hydroxypropyl-CoM dehydrogenase, and 2-oxopropyl-

CoM reductase) were absent in all class "Anaeropigmentia" genomes. Therefore, additional research is required to confirm the expression of CoM biosynthesis genes, and further characterize its potential function, if any, in class "Anaeropigmentia".

**Pigmentation:** The majority of Class "Anaeropigmentia" genomes encode *crtB*, 15-cis-phytoene synthase, and *crtI*, phytoene desaturase [EC:1.3.99.26 1.3.99.28 1.3.99.29 1.3.99.31], suggesting the capability of biosynthesis of lycopene from geranylgeranyl-PP. The gene encoding CrtY/L, lycopene beta-cyclase [EC:5.5.1.19], was however missing from all genomes, suggesting an acyclic carotenoid structure (Figure 3B).

**Class "Zymogenia"**

**General genomic features:** Genomes belonging to class "Zymogenia" possess average sized genomes (3.7 ± 0.12 Mbp), GC content (54.4% ± 2.7%), and gene length (904.24 ± 62.85 bp) (Table 1). Members of the class "Zymogenia" are Gram-negative rods with CRISPR and Type I restriction endonucleases as defense mechanisms and no intracellular microcompartments (Table S3).

**Physiological features:** Class "Zymogenia" genomes encode the capability for the compatible solute trehalose biosynthesis both from ADP-glucose and glucose via trehalose synthase, as well as from UDP-glucose and glucose via the action of trehalose 6-phosphate synthase and trehalose 6-phosphate phosphatase). No genes encoding trehalose degradation were identified in Class "Zymogenia" genomes. Biosynthesis and degradation of the storage molecule starch was encoded in the majority of Class "Zymogenia" genomes. Multiple genes encoding oxidative stress enzymes (catalase, rubrerythrin, rubredoxin, and alkylhydroperoxide reductase, peroxidase, and superoxide reductase) were identified (Table S3).

**Heterotrophic fermentative capacities:** Class "Zymogenia" genomes encode a glycolytic pathway, and a partial TCA pathway, suggesting heterotrophic capacities, possibly on a broad repertoire of sugars including glucose, fructose, galactose, lyxose, arabinose, sorbitol, and xylitol (Figure 4A, Table S3). Surprisingly, genes for fucose degradation to L-lactaldehyde (including L-fucose/D-arabinose isomerase, L-fuculokinase, and L-fuculose-phosphate aldolase) were encoded in the majority of Class "Zymogenia" genomes, but genes encoding the subsequent conversion of L-lactaldehyde to propanediol (L-lactaldehyde reductase), as well as genes encoding propanediol utilization (propanediol dehydratase) were missing. Genomes did not encode genes suggestive of aerobic (absence of complex III/ alternate complex III-encoding genes or anaerobic respiratory capacities, but encoded pyruvate fermentation genes. These include formate C-acetyltransferase and its activating enzyme catalyzing pyruvate fermentation to formate and acetyl-CoA, pyruvate ferredoxin oxidoreductase catalyzing pyruvate oxidative decarboxylation to acetyl-CoA, followed by conversion of acetyl-CoA to acetate with concomitant substrate level phosphorylation via the acetate CoA ligase, and acetolactate synthase and acetolactate decarboxylase for pyruvate conversion to (R)-acetoin (Table S3, Figure 4A). The lack of a complete electron transport chain, or genes encoding for anaerobic respiratory processes, argue for a predominantly fermentative lifestyle.

**ATP production and electron carrier recycling:** Beside substrate level phosphorylation, ATP production is also possible via the PMF-utilizing $H^+/Na^+$-transporting ATPase encoded in all genomes. All Class "Zymogenia" genomes encoded a complete WL pathway, and the majority encoded an RNF complex. WL pathway in Class "Zymogenia" is predicted to function as an electron sink, and the membrane-bound RNF complex is predicted to help in the generation of a proton motive force across the inner membrane while re-oxidizing the reduced ferredoxin produced from the action of pyruvate ferredoxin oxidoreductase. The PMF generated could be used for ATP production via the complete $H^+/Na^+$-transporting ATPase. The cytoplasmic electron bifurcating mechanism (HydABC) plus the heterodisulfide reductase MvhAGD-HdrABC (also encoded in the majority of genomes) would function to recycle electron carriers in the cytoplasm (Figure 4A).

### 4.4.3 Ecological Distribution

#### Class "Anaeroferrophillalia"

Fifty- four and 110 16S rRNA gene sequences affiliated with the class "Anaeroferrophillalia were identified in the IMG (Figure 2C) and NCBI nt (Figure S1) databases,
respectively (Nov. 2020). While "Anaeroferrophillalia" genomes were recovered from a limited number of locations (marine sediments, hydrothermal vents, thermal spring, and Zodletone spring), 16S rRNA analysis expanded their ecological distribution to a range of terrestrial (primarily wetlands and hydrocarbon-impacted environments), marine (predominantly hydrothermal vents, but also coastal and marine sediments), and freshwater (temperate, ground and thermal springs) environments (Figures 2C, S1B-G). The observed distribution patterns reinforce the metabolically predicted preference of members of the "Anaeroferrophillalia" to hypoxic and anoxic settings, as evident by preferences to the oxygen-poor wetlands and hydrocarbon-impacted habitats over grassland and agricultural soils in terrestrial settings, and the preferences to vents and marine sediments over pelagic samples in marine settings. Nevertheless, given the low number of total sequence and the extremely low percentage relative abundance, it is clear that members of the class "Anaeroferrophillalia" are perpetual members of the rare biosphere, and are rarely abundant community members in ecosystems analyzed to date.

#### Class "Anaeropigmentia"

Analysis of ecological distribution pattern identified 134 and 89 16S rRNA gene sequences affiliated with the class "Anaeropigmentia" in IMG (Figures 3C, S1B-G), and NCBI nt (Figure S1A) databases, respectively. While examined genomes were predominantly recovered from hypersaline environments (Ace Lake in Antarctica and Little Sippewissett salt marsh, MA, USA), the majority of 16S rRNA sequences associated with this group were largely associated with non-hypersaline freshwater temperate lake environments, e.g. Yellowstone Lake, the gas-saturated Lake Kivu, and a methane-emitting lake at the University of Notre Dame. Terrestrial environments harboring members of class "Anaeropigmentia" were predominantly wetlands, with hydrocarbon-impacted habitats being the only other terrestrial setting. A limited presence in coastal marine setting and absence from marine pelagic environments was observed. Notably, many of the environments harboring members

of class "Anaeropigmentia" are light-exposed (e.g. wetland surface sediment and lake water), justifying pigmentation, although many were not (e.g. coal mine soil and deep lake sediment). Similar to class "Anaeroferrophillalia", the limited number of affiliated 16S rRNA gene sequences suggests that members of class "Anaeropigmentia" are also part of the rare biosphere, present in small numbers across a range of different habitats.

### Class "Zymogenia"

Ecological distribution analysis identified only 44 and 33 16S rRNA gene sequences affiliated with the class "Zymogenia" in IMG (Figures 4B), and NCBI nt databases (Figure S1), respectively. A relatively high proportion of class "Zymogenia" 16S rRNA genes were recovered from oxygen-deficient environments, e.g. wetlands, marine sediments, and coastal sediments, as well as aquatic hypersaline settings, e.g. hypersaline lakes (Salton Sea) in California, and lagoons (Etoliko Lagoon in Greece). As well, a significant fraction of these sequences was recovered from anaerobic digestors and bioreactors environments, attesting to the preference of these organisms to anaerobic settings and adaptability of some its members to hypersaline environments (Figures 4B, S1B-G).

## 4.5 Discussion

Genomic analysis for members of class "Anaeroferrophillalia" revealed the capability of heterotrophic growth on a limited number of substrates, either fermentatively or using sulfur-cycle intermediates (polysulfide, thiosulfate, and tetrathionate) as electron acceptors. Autotrophic growth using the WL pathway and utilization of $H_2$ or Fe(II) as electron donors for chemolithotrophic growth in absence of organic carbon sources could also be inferred. Analysis of ecological distribution patterns identified the occurrence of class "Anaeroferrophillalia" as a rare component in few, mostly anaerobic, habitats. Such limited distribution could be a reflection of the limited range of substrates supporting its growth, as well as its dependence on the sulfur-cycle intermediates thiosulfate and tetrathionate as electron acceptors, rather than the more abundant, stable, and ubiquitous sulfate. Although uncommon, microorganisms depending on specific sulfur intermediates (thiosulfate, S, sulfite, tetrathionate) for growth, but not sulfate, for growth have previously been reported [244, 362]. Such pattern could be a reflection of metabolic interdependencies between various members of the sulfur cycle, a concept formulated based on the identification of a wide range of incomplete pathways in sequenced MAGs [10, 174]. In addition, previous studies have shown that thiosulfate and other sulfur cycle intermediates [56, 317], as well as iron [122], were present in much higher levels, and played a more pronounced role in supporting microbial growth on earth during geological eons preceding the evolution of oxygenic photosynthesis. Oxygen production and accumulation from photosynthesis has led to the slow but inexorable oxidation of earth's surface (great oxidation event), and the establishment of sulfate as the predominant electron acceptor in the sulfur cycle [56]. As such, rare lineages depending on $H_2$, Fe, and sulfur-intermediates (but not sulfate) for growth could represent lineages that thrived in a preoxygenated earth, but are rare nowadays.

Genomic analysis for members of Class "Anaeropigmentia" revealed a heterotrophic group of microorganisms capable of growing on sugars, amino acids, and fatty acids. Pref-

erence appears to be for anaerobic habitats, where it grows fermentatively, with predicted ability to grow under microaerophilic conditions. Genomic analysis also predicted the capacity for carotenoid (Lycopene) production. Carotenoid pigments are known to be present in photosynthetic bacteria, where they increase the efficiency of photosynthesis by absorbing in the blue-green region then transferring the absorbed energy to the light-harvesting pigments [153]. They also occur in a wide range of non-photosynthetic organisms, including Desulfobacterota, where they serve different purposes, including protection against desiccation [107], radiation [214], and oxidation [121]. The occurrence of members of the class "Anaeropigmentia" in transiently and intermittently light-exposed habitats, e.g. lakes and wetlands, could justify their pigmentation, as well as their capability to detoxify trace amounts of oxygen via respiration. [248, 268]

Finally, our genomic and ecological distribution pattern analysis for members of class "Zymogenia" predicted predominantly anaerobic fermentative organisms; and that some members of this novel class are adapted to growth under saline settings. We also note its extreme rarity in most examined settings. The underlying reason for such rarity is unclear, given its relatively wide substrate utilization range.

In conclusion, our work expands the metabolic and phylogenetic diversity of the Desulfobacterota through the description of 3 novel classes. Our analysis adds to the repertoire of ecological distribution and metabolic capabilities known for the phylum Desulfobacterota, with notable metabolic findings of iron metabolism, thiosulfate and tetrathionate reduction, carotenoid biosynthesis, CoM biosynthesis, and fermentation. Ecological distribution patterns observed reinforced and added context to the predicted metabolic capacities gleaned from genomic analysis. The study, overall, demonstrates the utility of bioinformatic tools in exploring and defining unculturable organisms, helping to bridge the vast knowledge gap presented by the uncultured majority.

## 4.6 Materials and Methods

### 4.6.1 Sample collection, DNA extraction, and metagenomic sequencing

Samples were collected from the source sediments of Zodletone Spring, located in western Oklahoma's Anadarko Basin (N34.99562° W98.68895°). The geochemistry of the spring has previously been described [47, 113, 347, 365]. Samples were obtained from the anoxic sulfidic black sediments at the source of the spring using sterile spatulas and were deposited into sterile 50 mL polypropylene plastic tubes. The samples were transferred to the laboratory on ice and immediately processed. DNA extraction was performed using the DNeasy PowerSoil kit (Qiagen, Valencia, CA, USA) according to manufacturer protocols. The extracted DNA was sequenced using the services of a commercial provider (Novogene, Beijing, China) using the Illumina HiSeq 2500 platform. 281.0 Gbp of raw data were obtained. Contigs were assembled using MegaHit [232], and binned using both Metabat [200] and MaxBin2 [405]. DasTool was used to select the highest quality bins from each metagenome assembly [353]. Bins that showed contamination levels >5% and/or strain heterogeneity of >10% were further refined and cleaned based on taxonomic affiliations of the bins, GC content, tetranucleotide frequency, and coverage levels using RefineM [291]. Bins were classified using the classification workflow option -classify_wf of GTDB-Tk [65] (v 1.3.0), and 5 bins belonging

to novel Desulfobacterota classes were selected for further analysis. In addition, we identified genomes belonging to the Desulfobacterota within the recently released 52,515 genomes in the earth microbiome catalogue collection [280] that were deposited in IMG/M database. Of these, 25 genomes belonging to novel Desulfobacterota classes were downloaded from the IMG/M database (May 2020) and included in the analysis.

### 4.6.2 Genomes quality assessment and general genomic features

Genome completeness, genome contamination, strain heterogeneity, and GC content were assessed using CheckM (v 1.0.13) [297]. Genomes with >50% completion and <10% contamination (n=30) were used for further analysis (Tables S1, S2). Selected MAGs were designated as medium or high-quality drafts based on the criteria set forth by MIMAGs [37]. The 5S, 16S, and 23S rRNA sequences were identified using RNAmmer (v 1.2) [223]. tRNA sequences were identified and enumerated with tRNAscan-SE (v 2.0.6, May 2020) using the -G general tRNA model [64].

### 4.6.3 Phylogenomic analysis

Preliminary classification was carried out using GTDB-Tk [65] with the -classify_wf option. Further phylogenomic analysis was conducted using the 120 single-copy marker genes concatenated alignment that was generated by GTDB-Tk [291]. A maximum-likelihood phylogenetic tree was constructed in RAxML (v 8.2.8) [369] with the PROTGAMMABLOSUM62 model and default settings, using members of the phylum Bdellovibrionota as an outgroup. Tree phylogeny, along with average amino acid identity (AAI), calculated using the AAI calculator ([http://enve-omics.ce.gatech.edu/]), were used to determine putative taxonomic ranks. The arbitrary AAI cutoffs used were 49%, 52%, 56%, and 68% for class, order, family, and genus, respectively [208, 298].

### 4.6.4 Functional annotation

Protein-coding genes were annotated using Prodigal (v 2.50) [179]. Identified protein-coding genes were assigned KEGG orthologies (KO) using BlastKOALA [197], and metabolic pathways were visualized with KEGG mapper [196]. For more targeted analysis of functions of interest, hidden markov model (HMM) profile scans were performed on individual genomes. All genomes were queried with custom-built HMM profiles for sulfur metabolism, and electron transport chain complexes. Custom profiles were built from Uniprot reference sequences for all genes with an assigned KO number, which were downloaded, aligned with Clustal-omega [354], and assembled into a profile with the hmmbuild function of HMMer (v 3.1b2) [258]. For genes without a designated KO number, a representative protein was queried against the KEGG genes database using Blastp, and hits with e-values $<1e^{-80}$ were downloaded, aligned, and used to construct an HMM profile as described above. HMM scans were carried out using the hmmscan function of HMMer [258]. A thresholding option of -T 100 was used to limit results to alignments with a score of at least 100 to improve specificity. Further confirmation was achieved through phylogenetic assessment and tree building procedures. Briefly, putatively identified sequences were aligned with Clustal-omega [354] against the reference sequences used to build the HMM database and placed into a maximum-likelihood

phylogenetic tree using FastTree (v 2.1.10) [309]. Sequences that clustered with reference sequences were deemed to be true hits and were assigned a corresponding KO number. FeGenie [130] was used to predict the presence of iron reduction and iron oxidation genes in individual bins. Hydrogenases were identified using HMM scans with profiles constructed from alignments from the Hydrogenase Database (HydDB) [376] using a cutoff e-value of $1e^{-20}$.

### 4.6.5 Ecological distribution

A near-complete 16S rRNA gene from each class was selected as a representative for querying against 16S rRNA databases. Representative sequences were queried against two different databases: 1. The IMG/M 16S rRNA publicly available assembled metagenomes [68], where an e-value threshold of $1e^{-10}$, percentage similarity $\geq$90%, and either $\geq$80% subject length for full-length query sequences or $\geq$80% query length for non-full-length query sequences criterion were applied, and 2. The GenBank nucleotide (nt) database (accessed November 2020), using a minimum identity threshold of 90%, $\geq$80% subject length alignment for near full-length query sequences or $\geq$80% query length for non-full-length query sequences, and a minimum alignment length of 100 bp. Hits meeting the selection criteria were then aligned with 16S rRNA reference gene sequences from each class using Clustal-Omega [354], and the alignment was used to construct maximum-likelihood phylogenetic trees with FastTree [309]. The environmental source of hits clustering with the appropriate reference sequences were then classified with a scheme based on the GOLD ecosystem classification scheme [266]. Phylogenetic trees were visualized and annotated in iTol [231].

### 4.6.6 Sequence and MAG accessions

Metagenomic raw reads for Zodletone sediment are available under SRA accession SRX9813571. Zodletone whole genome shotgun project was submitted to GenBank under Bioproject ID PRJNA690107 and Biosample ID SAMN17269717. The individual assembled MAGs have been deposited at DDBJ/ENA/GenBank under the accession JAFGAM000000000, JAFGAS000000000, JAFGFE000000000, JAFGIX000000000, JAFGSY000000000. The version described in this paper is version JAFGAM010000000, JAFGAS010000000, JAFGFE010000000, JAFGIX010000000, JAFGSY010000000.

## 4.7 Acknowledgements

This work has been supported by NSF grant 2016423 to NHY and MSE.

This work is published and available in full online (DOI: 10.1111/1462-2920.15614).

## 4.8 Figures & Tables

Supplementary Figures and Tables can be viewed online at:
https://sfamjournals.onlinelibrary.wiley.com/doi/10.1111/1462-2920.15614


Figure 4.1: Maximum likelihood phylogenetic tree based on: **(A)** a concatenated alignment of 120 single-copy genes from all Desulfobacterota classes in GTDB r95, and **(B)** 16S rRNA gene for Desulfobacterota classes with cultured representatives in GTDB r95. The three novel classes described here are colour-coded as shown in the legend. Bootstrap values (from 100 bootstraps) are displayed for branches with ≥70% support. Members of the phylum Bdellovibrionota were used as an outgroup (not shown).

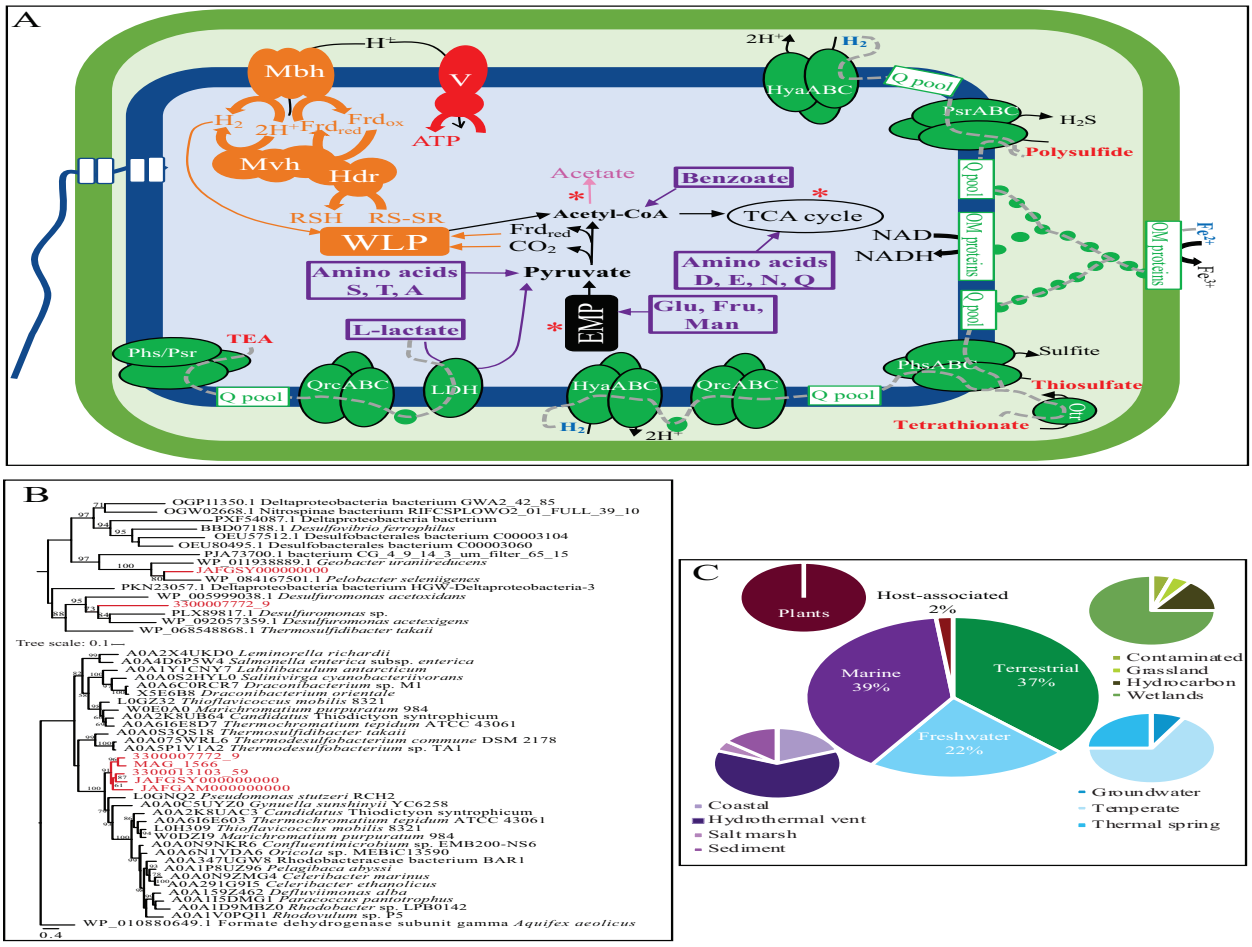



Figure 4.2: **A.** Cellular metabolic reconstruction based on genomic analysis of seven genomes belonging to the novel class *Candidatus* 'Anaeroferrophillalia'. Substrates predicted to support growth are shown in purple boxes, electron donors are shown in blue while electron acceptors are shown in red. Fermentation end products are shown in pink. Sites of substrate-level phosphorylation are shown as red asterisks. All electron transport chain components in the membrane are shown in green, while components for proton motive force creation and electron carrier recycling are shown in orange. Grey dotted lines depict the predicted flow of electrons from electron donors to electron acceptors. Green circles in the periplasmic space depict cytochromes. Abbreviations and gene names: EMP, Embden–Meyerhof–Paranas pathway; $Frd_{ox/red}$, Ferredoxin (oxidized/reduced); Fru, fructose; Glu, glucose; Hdr, heterodisulfide reductase complex; HyaABC, periplasmic [Ni Fe] hydrogenase; IM proteins, inner membrane protein complex for the predicted iron oxidation system; LDH, L-lactate dehydrogenase; Man, mannose; Mbh, membrane-bound [Ni Fe] hydrogenase; Mvh, Cytoplasmic [Ni Fe] hydrogenase; OM proteins, outer membrane protein complex for the predicted iron oxidation system; Otr, octaheme tetrathionate reductase; PhsABC, thiosulfate reductase; PsrABC, polysulfide reductase; QrcABC, menaquinone reductase; Q pool, quinone pool; RSH/RS-SR, reduced/oxidized disulfide; TCA, tricarboxylic acid cycle; TEA, terminal electron acceptor; V, ATP synthase complex; WLP, Wood Ljungdahl pathway. **B.** Phylogenetic affiliation for *Candidatus* 'Anaeroferrophillalia' thiosulfate reductase C subunit (PhsC, top) and the iron oxidation complex protein DFE_0462 (bottom) in relation to reference sequences. *Candidatus* 'Anaeroferrophillalia' sequences are shown in red. Alignments were created in Mafft [278] and maximum likelihood trees were constructed in RaxML [369]. Bootstrap support values are based on 100 replicates and are shown for nodes with ¿50% support. **C.** Ecological distribution of *Candidatus* 'Anaeroferrophillalia'-affiliated 16S rRNA sequences. The middle pie chart shows the breakdown of hit sequences based on the classification of the environments from which they were obtained (classification is based on the GOLD ecosystem classification scheme). Further sub-classifications for each environment are shown as smaller pie charts.

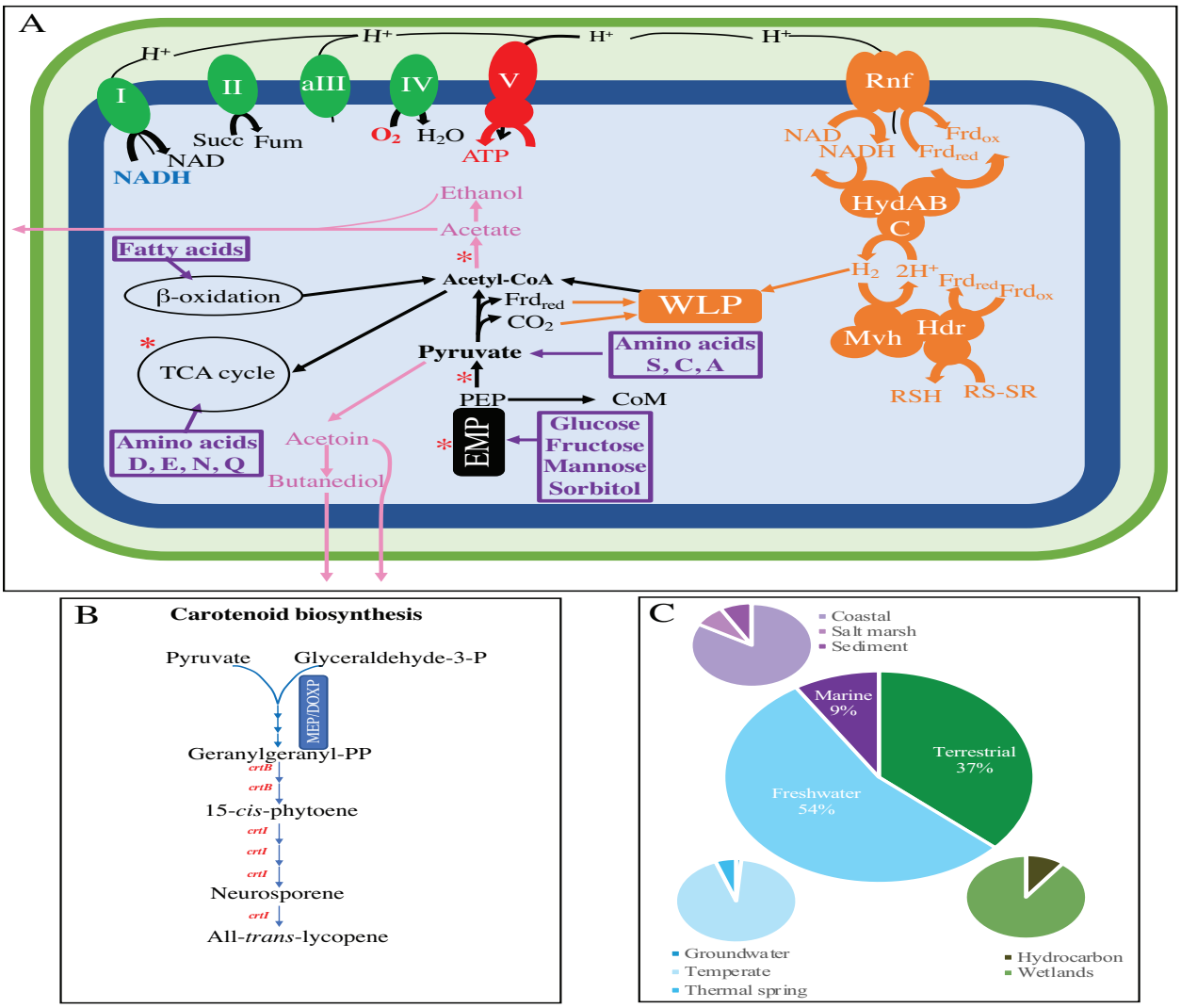



Figure 4.3: **A.** Cellular metabolic reconstruction based on genomic analysis of 17 genomes belonging to the novel class *Candidatus* 'Anaeropigmentia'. Substrates predicted to support growth are shown in purple boxes, electron donors are shown in blue while electron acceptors are shown in red. Fermentation end products are shown in pink. Sites of substrate-level phosphorylation are shown as red asterisks. All electron transport chain components in the membrane are shown in green, while components for proton motive force creation and electron carrier recycling are shown in orange. Abbreviations and gene names: CoM, coenzyme M; EMP, Embden–Meyerhof–Paranas pathway; $Frd_{ox/red}$, Ferredoxin (oxidized/ reduced); fum, fumarate; Hdr, heterodisulfide reductase complex; HydABC, cytoplasmic [Fe Fe] hydrogenase; I, II, aIII, and IV, aerobic respiratory chain comprising complexes I, II, alternate complex III, and complex IV; Mvh, Cytoplasmic [Ni Fe] hydrogenase; PEP, phosphoenol pyruvate; RNF, membrane-bound RNF complex; RSH/RS-SR, reduced/oxidized disulfide; succ, succinate; TCA, tricarboxylic acid cycle; V, ATP synthase complex; WLP, Wood Ljungdahl pathway. **B.** Carotenoid biosynthesis genes encountered in *Candidatus* 'Anaeropigmentia' genomes. Genes identified in at least one genome are shown in red boldface text, while genes with no homologues in the genomes are shown in black text. MEP/DOXP, the non-mevalonate DOXP/MEP (Deoxyxylulose 5-Phosphate/Methylerythritol 4-Phosphate) pathway for isoprenoid unit biosynthesis. **C.** Ecological distribution of *Candidatus* 'Anaeropigmentia'-affiliated 16S rRNA sequences. The middle pie chart shows the breakdown of hit sequences based on the classification of the environments from which they were obtained (classification is based on the GOLD ecosystem classification scheme). Further subclassifications for each environment are shown as smaller pie charts.

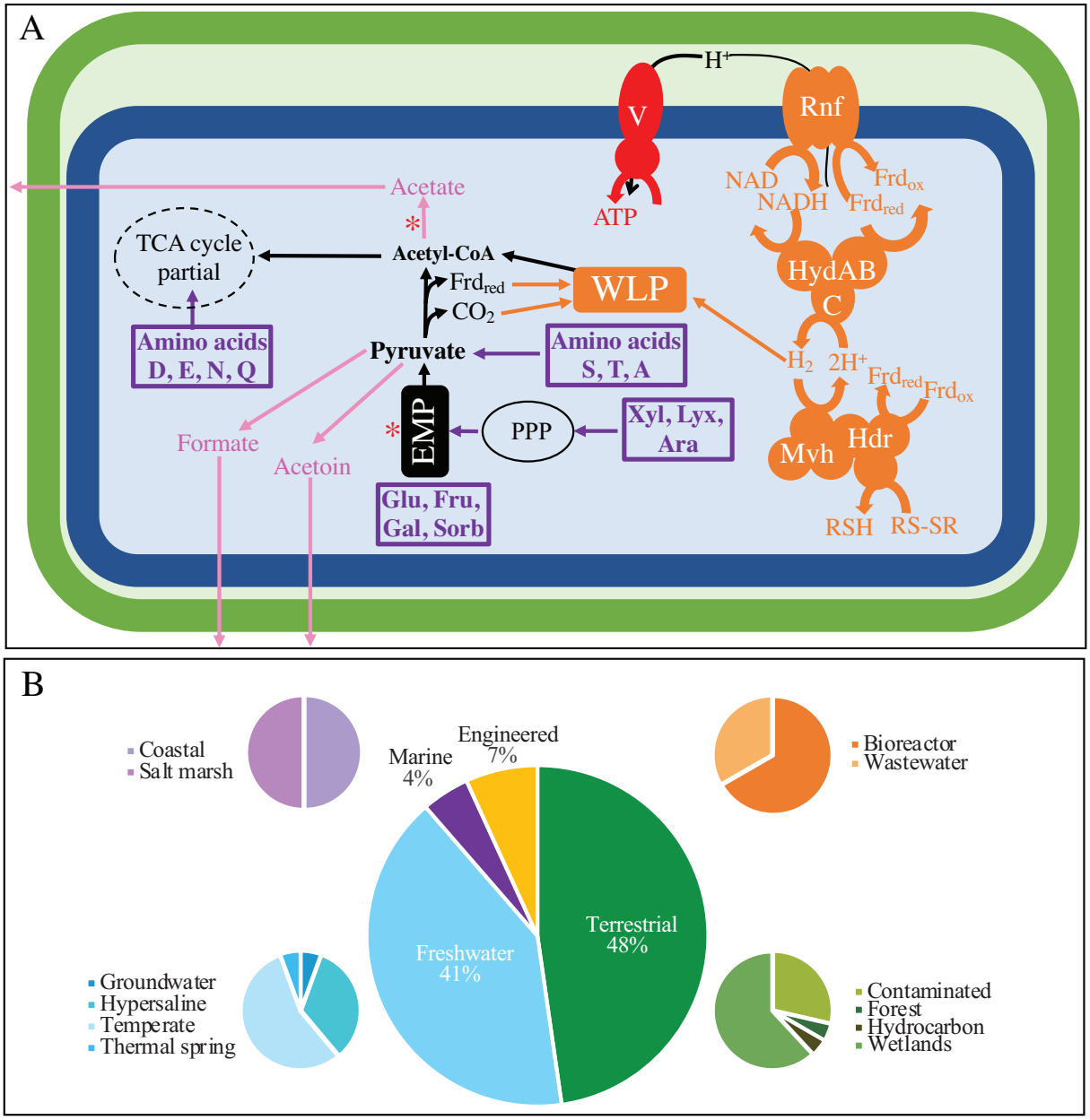



Figure 4.4: **A.** Cellular metabolic reconstruction based on genomic analysis of six genomes belonging to the novel class *Candidatus* 'Zymogenia'. Substrates predicted to support growth are shown in purple boxes, fermentation end products are shown in pink and sites of substrate-level phosphorylation are shown as red asterisks. Components for proton motive force creation and electron carrier recycling are shown in orange. Abbreviations and gene names: Ara, arabinose; EMP, Embden–Meyerhof–Paranas pathway; $Frd_{ox/red}$, Ferredoxin (oxidized/ reduced); Fru, fructose; Gal, galactose; Glu, glucose; Hdr, heterodisulfide reductase complex; HydABC, cytoplasmic [Fe Fe] hydrogenase; Lyx, lyxose; Mvh, Cytoplasmic [Ni Fe] hydrogenase; PPP, pentose phosphate pathway; RNF, membrane-bound RNF complex; RSH/RS-SR, reduced/oxidized disulfide; Sorb, sorbitol; TCA, tricarboxylic acid cycle; V, ATP synthase complex; WLP, Wood Ljungdahl pathway; Xyl, xylitol. **B.** Ecological distribution of *Candidatus* 'Zymogenia'-affiliated 16S rRNA sequences. The middle pie chart shows the breakdown of hit sequences based on the classification of the environments from which they were obtained (classification is based on the GOLD ecosystem classification scheme). Further sub-classifications for each environment are shown as smaller pie charts.

Table 1. General genomic features of the studied genomes.

| Bin Name† | GenBank assembly accession number† | Assembly Size (Mbp) | Expected Genome Size (Mbp) | GC % | Number of Genes | Coding % | Ecosystem classification (level 1) | Ecosystem sub classification (level 2) | Site Geographical Location | References | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | NCBI BioProject ID | Citation |
| 3300007772_9 | | 1.62 | 2.44 | 48.18 | 1653 | 91.58 | Marine | Hydrothermal vent | Axial Seamount | PRJEB22514 | 78 |
| 3300021512_13 | | 1.54 | 2.77 | 47.98 | 1543 | 89.89 | Marine | Hydrothermal vent | Guaymas Basin vent | PRJNA444987 | |
| 3300022547 | | 2.62 | 2.76 | 54.41 | 2438 | 88.24 | Freshwater | Thermal spring | Wilbur Springs, CA | PRJNA364781 | |
| 3300013103_59 | | 1.56 | 2.91 | 53.95 | 1523 | 89.96 | Marine | Hydrothermal vent | Guaymas Basin vent | PRJNA368391 | 79 |
| 3300031657_12 | | 1.82 | 2.53 | 47.22 | 1685 | 85.98 | Marine | Sediment | Southern Ocean, Antarctica | PRJNA518182 | |
| **Zgenome_940** | **JAFGSY000000000** | **2.7** | **2.74** | **53.55** | **2471** | **89.72** | **Freshwater** | **Sediment** | **Zodletone Spring, OK** | **PRJNA690107** | **This study** |
| Zgenome_0214 | JAFGAM000000000 | 2.68 | 3.44 | 63.31 | 2363 | 86.91 | Freshwater | Sediment | Zodletone Spring, OK | PRJNA690107 | This study |
| 3300023233_3 | | 4.81 | 5.07 | 41.66 | 4638 | 83.06 | Freshwater | Hypersaline | Ace Lake, Antarctica | PRJNA467265 | 80 |
| 3300023244_4 | | 3.96 | 4.39 | 41.42 | 3755 | 83.06 | Freshwater | Hypersaline | Ace Lake, Antarctica | PRJNA467197 | 80 |
| 3300023249_5 | | 4.15 | 4.49 | 41.9 | 3893 | 83.31 | Freshwater | Hypersaline | Ace Lake, Antarctica | PRJNA467248 | 80 |
| 3300025586_9 | | 3.77 | 4.56 | 42.11 | 3529 | 82.98 | Freshwater | Hypersaline | Ace Lake, Antarctica | PRJNA375651 | 80 |
| 3300025628_3 | | 4.04 | 4.64 | 41.98 | 3788 | 82.85 | Freshwater | Hypersaline | Ace Lake, Antarctica | PRJNA367219 | 80 |
| 3300025642_12 | | 3.83 | 4.7 | 42.16 | 3578 | 83.09 | Freshwater | Hypersaline | Ace Lake, Antarctica | PRJNA367224 | 80 |
| 3300028920_41 | | 3.47 | 4.6 | 50.24 | 3129 | 87.61 | Marine | Salt marsh | Falmouth, MA | PRJNA518321 | |
| **3300022855_4** | | **3.31** | **3.4** | **48.79** | **2957** | **85.77** | **Freshwater** | **Hypersaline** | **Ace Lake, Antarctica** | **PRJNA467247** | **80** |
| 3300022860_4 | | 3.34 | 3.68 | 49.14 | 3112 | 86.06 | Freshwater | Hypersaline | Ace Lake, Antarctica | PRJNA467267 | 80 |
| 3300023234_5 | | 3.18 | 3.35 | 49.09 | 2914 | 85.94 | Freshwater | Hypersaline | Ace Lake, Antarctica | PRJNA467206 | 80 |
| 3300023298_5 | | 3.22 | 3.39 | 48.95 | 2899 | 85.99 | Freshwater | Hypersaline | Ace Lake, Antarctica | PRJNA467209 | 80 |
| 3300025642_15 | | 3.19 | 3.36 | 48.98 | 2962 | 86.37 | Freshwater | Hypersaline | Ace Lake, Antarctica | PRJNA367224 | 80 |
| 3300025649_7 | | 3.79 | 3.92 | 48.75 | 3447 | 85.81 | Freshwater | Hypersaline | Ace Lake, Antarctica | PRJNA367218 | 80 |
| 3300028920_72 | | 2.23 | 3.25 | 57.19 | 2212 | 87.31 | Marine | Salt marsh | Falmouth, MA | PRJNA518321 | |
| 3300013123_31 | | 3.2 | 4.73 | 50.1 | 3098 | 84.18 | Freshwater | Temperate | Lake Kivu, Congo | PRJNA404433 | |
| 3300027902_63 | | 1.26 | 2.41 | 50.99 | 1326 | 88.93 | Freshwater | Sediment | University of Notre Dame, IN | PRJNA365732 | |
| 3300027515_27 | | 2.72 | 3.34 | 50.49 | 2445 | 89.96 | Freshwater | Thermal spring | Thermal spring, YNP | PRJNA367150 | |
| 3300017987_27 | | 3.56 | 3.68 | 52.62 | 3284 | 85.75 | Freshwater | Hypersaline | Salton Sea, CA | PRJNA444013 | |
| 3300017989_30 | | 3.48 | 3.66 | 52.53 | 3203 | 85.76 | Freshwater | Hypersaline | Salton Sea, CA | PRJNA444014 | |
| **Zgenome_24** | **JAFGIX000000000** | **3.76** | **3.84** | **51.37** | **3350** | **84.54** | **Freshwater** | **Sediment** | **Zodletone Spring, OK** | **PRJNA690107** | This study |
| 3300021603_37 | | 2.2 | 3.54 | 58.54 | 2397 | 87.12 | Engineered | Bioreactor | Toronto, ON | PRJNA501900 | |
| Zgenome_0311 | JAFGAS000000000 | 2.94 | 3.65 | 55.99 | 2730 | 86.46 | Freshwater | Sediment | Zodletone Spring, OK | PRJNA690107 | This study |
| Zgenome_1292 | JAFGFE000000000 | 3.73 | 3.84 | 55.37 | 3398 | 87.75 | Freshwater | Sediment | Zodletone Spring, OK | PRJNA690107 | This study |

†: Accession numbers here include the Metagenome Bin name (numbers starting with 33000; searchable at https://img.jgi.doe.gov/cgi-bin/mer/main.cgi?section=MetagenomeBinSearch&page=searchForm), and the NCBI assembly accession numbers for Zodletone MAGs.
Type strains are listed in **bold**.

Table 4.1: General genomic features of the studied genomes

# CHAPTER V

# GENOMES OF NOVEL MYXOCOCCOTA REVEAL SEVERELY CURTAILED MACHINERIES FOR PREDATION AND CELLULAR DIFFERENTIATION

## 5.1 Abstract

Cultured Myxococcota are predominantly aerobic soil inhabitants, characterized by their highly coordinated predation and cellular differentiation capacities. Little is currently known regarding yet-uncultured Myxococcota from anaerobic, non-soil habitats. We analyzed genomes representing one novel order (o__JAFGXQ01) and one novel family (f__JAFGIB01) in the Myxococcota from an anoxic freshwater spring (Zodletone spring) in Oklahoma, USA. Compared to their soil counterparts, anaerobic Myxococcota possess smaller genomes, and a smaller number of genes encoding biosynthetic gene clusters (BGCs), peptidases, one- and two-component signal transduction systems, and transcriptional regulators. Detailed analysis of thirteen distinct pathways/processes crucial to predation and cellular differentiation revealed severely curtailed machineries, with the notable absence of homologs for key transcription factors (e.g. FruA and MrpC), outer membrane exchange receptor (TraA), and the majority of sporulation-specific and A-motility-specific genes. Further, machine-learning approaches based on a set of 634 genes informative of social lifestyle predicted a non-social behavior for Zodletone Myxococcota. Metabolically, Zodletone Myxococcota genomes lacked aerobic respiratory capacities, but encoded genes suggestive of fermentation, dissimilatory nitrite reduction, and dissimilatory sulfate-reduction (in f_JAFGIB01) for energy acquisition. We propose that predation and cellular differentiation represent a niche adaptation strategy that evolved circa 500 Mya in response to the rise of soil as a distinct habitat on earth.

## 5.2 Importance

The Myxococcota is a phylogenetically coherent bacterial lineage that exhibits unique social traits. Cultured Myxococcota are predominantly aerobic soil-dwelling microorganisms that are capable of predation and fruiting body formation. However, multiple yet-uncultured lineages within the Myxococcota have been encountered in a wide range of non-soil, predominantly anaerobic habitats; and the metabolic capabilities, physiological preferences, and capacity of social behavior of such lineages remain unclear. Here, we analyzed genomes recovered from a metagenomic analysis of an anoxic freshwater spring in Oklahoma, USA that represent novel, yet-uncultured, orders and families in the Myxococcota. The genomes appear to lack the characteristic hallmarks for social behavior encountered in Myxococcota genomes, and displayed a significantly smaller genome size and a smaller number of genes

encoding biosynthetic gene clusters, peptidases, signal transduction systems, and transcriptional regulators. Such perceived lack of social capacity was confirmed through detailed comparative genomic analysis of thirteen pathways associated with Myxococcota social behavior, as well as the implementation of machine learning approaches to predict social behavior based on genome composition. Metabolically, these novel Myxococcota are predicted to be strict anaerobes, utilizing fermentation, nitrate reduction, and dissimilarity sulfate reduction for energy acquisition. Our results highlight the broad patterns of metabolic diversity within the yet-uncultured Myxococcota and suggest that the evolution of predation and fruiting body formation in the Myxococcota has occurred in response to soil formation as a distinct habitat on earth.

## 5.3 Introduction

The "Myxobacteria" represent a phylogenetically coherent lineage within the Domain Bacteria [351]. Originally assigned to the class deltaproteobacteria [350, 366], they have recently been recognized as a separate phylum (Myxococcota) based on phylogenomic assessment; a proposal empirically supported by their distinct metabolic, and structural characteristics [390]. The Myxococcota are highly social organisms, displaying specific behaviors (predation and fruiting body formation) that require a high level of kin recognition, cell-to-cell communication, and intercellular coordination [58]. Indeed, cellular differentiation in the Myxococcota has aptly been described as the most successful foray for a prokaryotic organism into multicellularity [58].

Both predation and fruiting body formation processes involve differential gene activation and expression in seemingly equivalent cells, leading to distinct cellular differentiation and disparate fates in response to external environmental stimuli. As predators, model Myxococcota utilize an epibiotic strategy, where swarms of motile cells surround and lyse prey cells via the production of secondary metabolites and extracellular enzymes [379]. Significant coordination of motility and lytic agents production between individual cells has been proposed as a means to enhance the efficiency of Myxococcota predator swarms ([27] but see [247]). Cellular differentiation in the Myxococcota entails the formation of elaborate multicellular structures (fruiting bodies) in response to nutrient depletion [272]. The process involves cell aggregation and subsequent differentiation of a subset of cells into resistant myxospores, another subset into peripheral rods on the outside of the fruiting body adapted to rapidly respond to re-appearing nutrients, and a third subset undergoing programmed cell death [271, 272].

An extensive body of literature, spanning decades on the mechanistic basis of social behavior in model Myxococcota, is available [59, 216, 272, 342, 379, 420]. Predation is enabled by two types of gliding motility (individual adventurous (A) motility and cooperative social (S) motility), exopolysaccharide production, secretion of secondary metabolites and extracellular enzymes (proteases and CAZymes), and mechanisms for kin/self-recognition and cheaters elimination [58] (this later is also important during aggregation and fruiting body formation). Starvation-induced cellular differentiation is mediated by four modules of highly interconnected and signal-responsive cascades of signal transduction networks that act sequentially as well as cooperatively to coordinate and time aggregation, fruiting body formation, and sporulation [216]. Simultaneously, starvation-induced stringent response prompts

the production of extracellular signals (most importantly A-signal, and C-signal) that activate a wide range of transcription factors to facilitate aggregation, regulate the onset of sporulation, and complete the development process [216].

Most cultured Myxococcota, henceforth referred to as "model Myxococcota", are aerobic soil dwellers, known to inhabit the top layers of agricultural, forest, and even desert soils [321]. This strong niche preference pattern attests to the contribution of their unique capacities to their success in soil ecosystems. The dual saprophytic/predatory capacities allow Myxococcota to utilize live microbial cells as well as microbial, floral, and faunal detritus as food sources [321]. Their gliding motility allows them not only to get in close proximity with their prey, but also to access their insoluble substrates in soils [321]. Their social behavior allows sharing resources (especially exoenzymes), enabling a more efficient process in which higher enzymatic activity is achieved when compared to individual cells [321]. Fruiting body formation guarantees long term survival under adverse and highly fluctuating conditions in soil, as well as faster recovery and propagation under more favorable circumstances.

However, while Myxococcota appears to be most successful and prevalent in soils, members of this phylum have also been isolated from non-soil habitats (e.g. Nannocystaceae and *Haliangiaceae* from aerobic marine sediments [131, 133], and the facultative anaerobic *Anaeromyxobacter* from contaminated soils and sediments [103]). Further, multiple amplicon-based diversity surveys have identified Myxococcota-affiliated sequences in non-soil habitats, many of which represent anoxic/ hypoxic settings [181, 188, 195, 210, 235, 262, 381]. Recently, the implementation of genome-resolved metagenomic approaches has resulted in the recovery of Myxococcota genomes from a wide range of non-soil habitats, almost invariably constituting a minor fraction of the population [7, 191, 283, 329, 337, 409]. Interestingly, many of these yet-uncultured lineages identified using 16S rRNA gene amplicon or metagenomic surveys appear to represent distinct novel, yet-uncultured lineages within the Myxococcota.

Given these observed strong niche adaptation patterns, as well as the predicted correlation between Myxococcota predation and cellular differentiation capacities, and successful propagation in soil, we hypothesized that novel, yet-uncultured Myxococcota recovered from anaerobic non-soil habitats will display distinct metabolic, physiological, and lifestyle capacities when compared to their soil counterparts. Here, we analyzed multiple genomes representing two novel Myxococcota lineages recovered from the completely anoxic, sulfide-laden (8-10 mM) source sediment in Zodletone spring, an anaerobic sulfide-and sulfur-rich spring in Oklahoma. We investigated the metabolic capacities, physiological preferences, structural features, and potential social behavior of these lineages, and compared their predicted capacities to model social aerobic Myxococcota. Our results suggest that non-soil Myxococcota possess severely curtailed pathways for the typical social behavior of soil Myxococcota, potentially utilize fermentation and/or sulfate-reduction for energy generation as opposed to aerobic respiration, and show preferences to polysaccharide and sugars, rather than proteins and amino acids as carbon and energy sources. We argue that such differences provide important clues to the evolution of social behavior in the Myxococcota in light of our understanding of the history of soil formation and oxygen accumulation in the atmosphere.

## 5.4 Materials and Methods

### 5.4.1 Site description and geochemistry

Zodletone spring is located in the Anadarko Basin of western Oklahoma (N34.99562° W98.68895°). The site geochemistry has been previously described in detail [47, 347, 364]. Briefly, at the spring source, sulfide and gaseous hydrocarbon-saturated waters are slowly (8 L/min) ejected, along with sediments that deposit at the source of the spring. High (8-10 mM) sulfide concentrations maintain complete anoxic conditions (oxygen levels $< 0.1$ µM) in the spring sediments. Oxygen concentrations in the 50 cm water column overlaying the sediments vary from 2–4 µM at the 2 mm above the source to complete oxygen exposure on the top of the water column [47].

### 5.4.2 Sampling and nucleic acid extraction

The sampling and DNA extraction processes have been previously described in detail [267, 407]. Briefly, ten different sediment samples (≈50 grams each) were collected at 5-cm depth, as well as from the standing overlaid water in sterile containers. DNA was extracted from 0.5 grams of source sediments from each replicate sample. For water samples, 10L of water was filtered on 0.2 $\mu$m sterile filters, and DNA was directly extracted from filters. Extraction was conducted using the DNeasy PowerSoil kit (Qiagen, Valencia, CA, USA) according to manufacturer protocols.

### 5.4.3 Metagenome sequencing, assembly, and binning

All extractions from sediment or water samples were pooled, and the pooled DNA was used for the preparation of sequencing libraries using the Nextera XT DNA library prep kit (Illumina, San Diego, CA, USA) as per manufacturer's instructions. DNA sequencing was conducted using two lanes on the Illumina HiSeq 2500 platform and 150-bp pair-end technology for each of the water and sediment samples using the services of a commercial provider (Novogene, Beijing, China). Metagenomic sequencing of the sediments and water samples yielded 281 Gbp and 323 Gbp of raw data, respectively. Reads were assessed for quality using FastQC followed by quality filtering and trimming using Trimmomatic v0.38 [33] using a sliding window size of 4 bases and a sliding window average minimum quality of 15, a leading and trailing minimum quality of 3, and a minimum read length of 36. High quality reads were assembled into contigs using MegaHit (v.1.1.3) [232] with minimum Kmer of 27, maximum kmer of 127, Kmer step of 10, and minimum contig length of 1000 bp. Bowtie2 was used to calculate the percentage of reads that assembled into contigs and sequencing coverage for each contig. Contigs >1 Kbp were binned into draft metagenome-assembled genomes (MAGs) using Metabat [200] and MaxBin2 [405], followed by selection of the highest quality bins using DasTool [353]. CheckM [297] was used for estimation of genome completeness, strain heterogeneity, and contamination by employing the lineage-specific workflow (lineage_wf flag). Quality designation of draft genomes was based on the criteria set forth by MIMAGs [37].

### 5.4.4 Genome classification

Taxonomic classifications followed the Genome Taxonomy Database (GTDB) release r95 [292], and were carried out using the classify_workflow in GTDB-Tk (v1.3.0) [65]. Phylogenomic analysis utilized the concatenated alignment of a set of 120 single-copy bacterial genes [292] generated by the GTDB-Tk. Maximum-likelihood phylogenomic tree was constructed in RaxML using the PROTGAMMABLOSUM62 model and default parameters [212] and members of the Bdellovibrionota as an outgroup. To further assign genomes to putative families and genera, average amino acid identity (AAI), and shared gene content (SGC) were calculated using the AAI calculator [http://enve-omics.ce.gatech.edu/]. The arbitrary AAI cutoffs used were 49%, 52%, 56%, and 68% for class, order, family, and genus, respectively [208, 298]. Further, Relative Evolutionary Divergence (RED) values, based on placement in the GTDB backbone tree (available at www.data.gtdb.ecogenomic.org/releases/release95/95.0/), were used to confirm the novelty of lineages to which the genomes are assigned. Values between 0.62 and 0.46 are indicative of a novel order, and values between 0.62 and 0.77 of a novel family.

### 5.4.5 Annotation and genomic analysis

Protein-coding genes were predicted using Prodigal [179]. GhostKOALA [197] was used for the annotation of every predicted open reading frame in bins and to assign protein-coding genes to KEGG orthologies (KOs), followed by metabolic pathways visualization in KEGG mapper [196]. In addition, all genomes were queried with custom-built HMM profiles for alternate complex III components and hydrogenases. To construct HMM profiles, a representative protein was queried against the KEGG genes database using Blastp, and hits with e-values $<1e^{-80}$ were downloaded, aligned, and used to construct an HMM profiles using the hmmbuild function of HMMer (v 3.1b2) [258]. Hydrogenases HMM profiles were built using alignments downloaded from the Hydrogenase Database (HydDB) [376]. The hmmscan function of HMMer [258] was used with the constructed profiles and a thresholding option of -T 100 to scan the protein-coding genes for possible hits. Further confirmation was achieved through phylogenetic assessment and tree building procedures. The 5S, 16S, and 23S rRNA sequences were identified using Barrnap 0.9 (https://github.com/tseemann/barrnap). tRNA sequences were identified using tRNAscan-SE (v 2.0.6, May 2020) [64]. Genomes were mined for CRISPR and Cas proteins using the CRISPR/CasFinder [83]. Proteases, peptidases, and protease inhibitors were identified using Blastp against the MEROPS database [318], while carbohydrate active enzymes (CAZymes) were identified by searching all ORFs from all genomes against the dbCAN hidden Markov models V9 [171] (downloaded from the dbCAN web server in September 2020). AntiSMASH 3.0 [252] was used with default parameters to predict biosynthetic gene clusters in the genomes. Metabolic reconstruction of reference Myxococcota type species genomes was obtained from the KEGG genomes database (https://www.genome.jp/kegg/genome/) and used for comparative genomics to Zodletone Myxococcota genomes.

### 5.4.6 Phylogenetic analysis of dissimilatory sulfite reductase DsrAB

Predicted dissimilatory sulfite reductase subunits A and B were compared to reference sequences for phylogenetic placement by first aligning to corresponding subunits from sulfate-reducing taxa using Mafft [278]. DsrA and DsrB alignments were concatenated in Mega X [221], and used to construct maximum-likelihood phylogenetic trees using FastTree (v 2.1.10) [309].

### 5.4.7 Machine learning approaches

Genomes of type species of cultured social (i.e. experimentally verified to be involved in predation and observed to undergo cellular differentiation and fruiting body formation) Myxococcota lineages (n=24) and all cultured non-social Myxococcota lineages (n=13) were downloaded from GenBank (June 2021). Lineages were assigned their "social" status using prior culture-based observations [103, 131, 132, 133]. Genomes were annotated with KO numbers using GhostKOALA [197] using default parameters, and gene counts were assembled into a matrix. Informative KOs (n=634) were selected using indicator analysis, with the R package indicspecies [93] using the multipattern function, and used to build a predictive model. Data was then centered, scaled, and Box-Cox transformed using the R package caret (https://topepo.github.io/caret/). Random forest classification training was performed in Python3 with the ensemble method of scikit-learn (v 0.24.1) [302]. The data was randomly divided, with 75% of the data selected to serve as the training set and the remaining 25% reserved for model verification. Model training was performed with default parameters and 1000 estimators. The model successfully predicted the social behavior (with 100% accuracy) in the 25% data subset reserved for model verification. Social abilities of novel Myxococcota lineages were then predicted using the constructed model. Matthew's Correlation Coefficient [70] was used to quantify classification accuracy. Including only pure cultures genomes' in our model with experimentally verified social behavior ensured that the model was accurately trained on possession of social behavior rather than arbitrary genomic artifacts due to phylogenetic relatedness.

### 5.4.8 Sequence and MAG accessions

The individual assembled Myxococcota MAGs analyzed in this study have been deposited at DDBJ/ENA/GenBank under the accessions JAFGVO000000000, JAFGQN000000000, JAFGWT000000000, JAFGTB000000000, JAFGXQ000000000, and JAFGIB000000000.

## 5.5 Results

### 5.5.1 Comparative genomic analysis between Zodletone and model Myxococcota

Comparative genomic analysis between Zodletone Myxococcota MAGs and genomes of all described type species in the phylum Myxococcota (n=27) was conducted. These genomes belong to classes Myxococcia (n=12), Polyangia (n=12), and Bradymonadia (n=3), and 20 of which exhibit the distinct social behavior of the Myxococcota [27, 58, 103, 131, 132,

133, 216, 247, 272, 342, 379, 420]. We utilized only genomes from type species to ensure the availability of experimental data regarding various aspects of their lifestyle. Compared to model Myxococcota genomes (i.e. those shown to exhibit predation behavior and to form fruiting bodies, n=20); Zodletone genomes were significantly smaller (6.15 ± 1.28 Mb versus 11.44 ±2.52 Mb), with a lower gene count (5129±1005 versus 9461±1611) and GC content (49.53± 5.67 versus 69.68± 1.74) (Student t-test p-value <0.00001) (Figure 2). Further, multiple additional differences were observed in gene families previously implicated in mediating Myxococcota social lifestyle between Zodletone MAGs and genomes of social Myxococcota. Extracellular proteolytic enzymes are crucial components of the predatory machinery in Myxococcota, aiding in degrading prey-released proteins and/or inducing prey lysis [379]. Zodletone genomes encoded a significantly lower number of proteases/peptidases when compared to model Myxococcota (58 ± 3.4 versus 130 ±17) (Figure 2, Table S2). Of note is the absence of representatives of MEROPS Family M15 (peptidoglycan endopeptidases) specifically implicated in prey cell lysis (Table S2). Further, model Myxococcota also secrete a plethora of secondary metabolites such as pigments, siderophores, bacteriocins, and antibiotics that attack and lyse their prey [379]. Zodletone Myxococcota genomes encoded a significantly lower number of biosynthetic gene clusters (BGCs, 8±3), mostly belonging to the NRPS_PKS type. By comparison, model Myxococcota encoded a larger number of BGCs (38 ± 16), belonging to a wider range (NRPS, PKS, terpenes, siderophores, and phenazines) of BGC classes. (Figure 2, Table S3).

Predation in model soil Myxococcota is also associated with secretion of extracellular or outer membrane CAZymes for targeting prey cell walls. While the overall numbers of CAZymes encoded in Zodletone genomes were not significantly different from those encountered in model soil Myxococcota genomes (Figure 2), the CAZyome families were significantly different between the two groups (Table S4). Specifically, model Myxococcota genomes were enriched in two GH families: GH23 (peptidoglycan lyases, consistent with their ability to target prey cell walls), and GH13 amylases (Student t-test p-value <0.02). Instead, Zodletone genomes were significantly (Student t-test p-value <0.02) enriched in GH and PL families targeting polysaccharide degradation, e.g. GH12, GH5, GH45, GH8, GH9 endoglucanases and cellobiohydrolases for cellulose degradation; GH10, GH11 xylanases for hemicellulose backbone degradation, and GH43 and GH54 xylosidases for hemicellulose side chain sugar removal; and PL1 and PL11 pectin/pectate/rhamnogalacturonan lyases for pectin degradation.

Finally, the collective social behavior in model soil Myxococcota is underpinned by an expanded arsenal of transcriptional factors. These include signal transduction one-component systems (OCS) (with a sensory domain and a response effector domain present in the same gene) and two-component systems (TCS) (with a sensor histidine kinase (HK), a partner response regulator (RR), and occasionally a phosphotransfer protein (PP)), as well as other transcriptional factors (TFs) including transcriptional regulators (TRs), and alternative sigma factors (SFs) [**398**]. Model soil Myxococcota genomes encode 241 ± 87 OCS genes, 329 ± 95 TCS genes, and 127 ± 56 TF genes. In contrast, Zodletone Myxococcota genomes encoded significantly lower OCSs, TCSs, and TFs (65±14, 198±58, and 69±18, respectively) (Student t-test p<0.05) (Figure 2, Table S5). This pattern of curtailed transcription factor repertoire in Zodletone genomes was pronounced in OCS and TCS systems (Figure 2), specifically OCS families AraC, ArsR, GntR, LysR, MarR, TetR, and Xre, and

TCS-RR belonging to the families CheY, NarL, OmpR, and FrzZ (Table S5) (Student t-test p-value <0.05).

### 5.5.2 Comparative genomics analysis of predation and cellular differentiation genes/pathways in the Myxococcota

We assessed the distribution patterns of pathways implicated in Myxococcota social behavior in Zodletone Myxococcota and compared them to *Myxococcus xanthus*, the model myxobacterium extensively studied for its social behavior, as well as to *Anaeromyxobacter dehalogenes*, *Vulgatibacter incomptus*, and *Labilithrix luteola*. These later three isolates share many genomic features with social Myxococcota (large genome size, high %GC, large number of genes), but lack predation and fruiting body formation capacities [344, 408]. Thirteen pathways were examined: four gene regulatory networks governing sporulation, aggregation, and fruiting body formation; exopolysaccharide production genes; two extracellular signals production gene clusters (A-signal and C-signal); aggregation, sporulation, and fruiting body formation genes; chemosensory pathways; developmental timers; two motility gene clusters (A-motility and S-motility); and outer membrane exchange genes (Figure 3, Tables 2, and S6, S. text). Detailed analysis of the distribution patterns of genes in these pathways, as well as a background on their known functions is presented in the supplementary document (S. text). Collectively, the analysis clearly demonstrates that social behavior pathways were severely curtailed in Zodletone Myxococcota genomes (Tables 2, and S6, S. text), where homologues of genes specific for the model Myxococcota social lifestyle (e.g. sporulation, extracellular signal production, motility, outer membrane exchange) were missing from Zodletone genomes. Specifically, the most notable deficiencies were: 1. Absence of homologues for extracellular signals production that control early events in aggregation (Tables 2, and S6, S. text). 2. Absence of homologues for the two transcription factors FruA and MrpC that work cooperatively to control the start of sporulation [259, 285, 331], although we acknowledge that this absence has also been noted outside the suborders *Cystobacterineae* [176]. 3. Absence of homologues for sporulation specific genes [41, 216, 259, 274, 331], A motility-specific genes, and the outer membrane exchange receptor TraA that recognizes kin and allow membrane fusion [342]. Such pattern was also observed in the genomes of *Anaeromyxobacter dehalogenes. Labilithrix luteola,* and *Vulgatibacter incomptus*, all of which have been experimentally shown to lack the capacity to aggregate into mounds, form fruiting bodies, or sporulate in pure culture. Further, for pathways with homologues identified in Zodletone genomes, the majority of such homologues encoded genes that are widely distributed in bacterial genomes and not specific to the Myxococcota, e.g. transcriptional response regulators, serine/ threonine kinases, peptidase domains, guanylate cyclase domains, chemotaxis-associated domains, or type IV pili.

### 5.5.3 Machine learning approaches suggest absence of social behavior in Zodletone Myxococcota

It is important to note that most of the physiological, mutational, and transcriptomic studies on soil Myxococcota were conducted on the model organism *M. xanthus*, (Class Myxococcia), a relatively distant relative of Zodletone Myxococcota (Class Polyangia). Further, while

genes for exopolysaccharide production, adventurous motility, extracellular signal production, and sporulation are highly specific, a large proportion of the gene regulatory network and developmental timing proteins governing aggregation, sporulation and fruiting body formation are homologues to signal transduction proteins involved in various cellular processes in a wide swath of lineages, and are hence universally distributed within the bacterial world. Similarly, chemosensory network proteins in Myxococcota are homologues to a wide range of chemotaxis proteins. Indeed, in the genomes of the non-social *Anaeromyxobacter dehalogenes, Labilithrix luteola,* and *Vulgatibacter incomptus*, homologues for the highly specific extracellular signal production, and sporulation genes were not identified, but homologues for chemosensory networks and gene regulatory networks were found, attesting to the above caveats with the approach.

Therefore, as a complementary approach, we employed a machine learning technique to predict social behavior potential in Zodletone Myxococcota. The approach depends on first identifying, from the initial set of all genes in the genomes, a group of genes with assigned KO numbers in the genomes of known social Myxococcota that are absent from genomes of known non-social Myxococcota. These candidate genes are then used for model training using Random Forest algorithm, and the constructed model is then employed to predict the social behavior based on the genomic content of Zodletone Myxococcota genomes. The occurrence pattern of the list of 634 KOs (Dataset 1) selected for model training predicted non-social behavior for Zodletone Myxococcota lineages with a Matthew's correlation coefficient of +1, confirming the patterns observed with the comparative genomics approach detailed above.

### 5.5.4 Structural features and metabolic capacities

Structurally, Zodletone Myxococcota MAGs encoded determinants of Gram-negative cell walls (LPS biosynthesis and $G^-$ peptidoglycan structure), motility (flagellar assembly and type IV pili), pigmentation (carotenoid biosynthesis), chemotaxis, and rod-shape (MreBCD and RodA). The genomes also encoded Type III (partial) and type VI secretion systems. Such characteristics are similar to those displayed by vegetative cells of cultured Myxococcota (Figure S1a, Table S7).

Genomic analysis predicted key differences in anabolic capacities between Zodletone Myxococcota and cultured Myxococcota. Zodletone MAGs did not encode the capacity for glycogen or trehalose biosynthesis, both of which are biosynthesized and used as storage molecules by cultured Myxococcota, and shown to be essential for sporulation [89]. Additionally, evidence for a glyoxylate shunt were missing from Zodletone MAGs. The glyoxylate shunt is employed by cultured Myxococcota to bypass $CO_2$ loss and NADH production during the TCA cycle and drive the metabolism towards oxaloacetate in preparation for gluconeogenesis [274]. Further, key differences in levels of amino acids auxotrophy were predicted, where Zodletone Myxococcota MAGs encoded capacities for biosynthesis of almost all amino acids, compared to the observed auxotrophy for branched chain amino acids (in 9 type species) and aromatic amino acids (in 5 type species) in cultured Myxococcota [43, 110]. Such pattern reflects the dependence of cultured Myxococcota on proteins and amino acids as substrates (and hence their ready availability for biosynthetic purposes) as opposed to the lack of such capacity in Zodletone genomes, necessitating amino acid biosynthesis from metabolic precursors. Finally, while cultured Myxococcota are able to incorporate sulfur from sulfate as well

as organic sources (e.g. taurine, alkane sulfonate, and dimethyl sulfone in 6 type species), and incorporate N from ammonia as well as organic sources (e.g. urea in 11 type species), such capacities for S or N incorporation from organic sources were not encoded in Zodletone Myxococcota MAGs (Table S7). Additionally, order JAFGXQ01 genomes encoded the capability to fix atmospheric nitrogen (Figure S1a, Table S7).

Genomic analysis also demonstrated multiple key differences in catabolic processes (substrate utilization patterns, respiratory capacities, electron recycling pathways, and ATP generation mechanisms) between Zodletone and model Myxococcota genomes. While most model Myxococcota (with the exception of *Sorangium cellulosum*) rely on amino acids and lipids as substrates and are poor carbohydrate consumers [89], Zodletone genomes encode a much lower number of amino acid degradation pathways (only 9, compared to 15 in type species), consistent with their observed limited proteolytic capabilities (Table S2). Instead, Zodletone Myxococcota appear to possess a more extensive carbohydrate degradation capacity (Tables 3, S7, Figure S1a), with pathways enabling the degradation of nine different sugars, sugar alcohols, sugar amines, and uronic acids encoded in their genomes. This is consistent with the possession of a wide range of polysaccharide-degrading CAZymes, as described above (Table S4). Finally, Zodletone order JAFGXQ01 genomes encoded an incomplete beta-oxidation pathway for long chain fatty acid degradation (Table 3, Figure S1a), a pathway commonly occurring in cultured Myxococcota to enable fatty acid consumption as the main carbon and energy source.

All cultured Myxococcota (with the exception of the genus *Anaeromyxobacter*) are aerobic microorganisms. In contrast, Zodletone Myxococcota genomes lack key genes for aerobic respiration, specifically homologues for either complex III or alternative complex III, as well as the absence of homologues for the low affinity cytochrome oxidase aa3. High affinity cytochrome bd ubiquinol oxidase is encoded in Zodletone genomes, but could possibly be employed in detoxification of trace amounts of $O_2$ present. Instead, the genomes encode genes enabling the utilization of nitrite as a terminal electron acceptor via the cytochrome-linked nitrite reductase NrfAH (Table 3, Figure S1a). Further, the single genome representative of novel family JAFGIB01 encodes a full dissimilatory sulfate reduction machinery including sulfate adenylyltransferase (Sat; EC 2.7.7.4) for sulfate activation to APS, adenylylsulfate reductase (AprAB; EC:1.8.99.2) for adenylyl sulfate reduction to sulfite, QmoABC for electron transfer, dissimilatory sulfite reductase (DsrAB; EC:1.8.99.5) and its co-substrate DsrC for dissimilatory sulfite reduction to sulfide, and the sulfite reduction-associated membrane complex DsrMKJOP for linking cytoplasmic sulfite reduction to energy conservation (Table 3, Figure S1a). JAFGIB01 genome also encoded octaheme tetrathionate reductase (*otr*) and thiosulfate reductase *phsABC,* suggesting the capability to utilize tetrathionate and thiosulfate as terminal electron acceptors in addition to sulfate (Figure S1a). This is the first genomic report of sulfur species respiration capability in the Myxococcota, and could possibly be a reflection of the sulfur and sulfide-rich Zodletone Spring environment from which the MAG was binned. Phylogenetically, JAFGIB01 DsrAB were most closely affiliated to DsrAB sequences encountered in Acidobacteria genomes (Figure S1b) [407].

Besides respiration, additional pathways for electron disposal were identified in Zodletone Myxococcota genomes. These include fermentative processes for acetate, ethanol, and lactate production from pyruvate (Table 3, Figure S1a). As well, the genomes encoded a full Wood Ljungdahl pathway (WLP), most probably acting as an electron sink mechanism

for re-oxidizing reduced ferredoxin, as previously noted in *Candidatus* Bipolaricaulota and Desulfobacterota genomes [267, 415]. Finally, a possible additional mechanism for ATP production in Zodletone Myxococcota is the utilization of the RNF complex for re-oxidizing reduced ferredoxin at the expense of NAD, with the concomitant export of protons to the periplasm, generating a proton motive force that can drive ATP production via oxidative phosphorylation via the encoded F-type ATP synthase. Consistent with encoding RNF complex components, the genomes also encoded elements for electron carriers recycling including the cytoplasmic electron bifurcating mechanism HydABC. Analysis of Myxococcota type species genomes revealed absence of RNF complex components, HydABC electron bifurcation system, as well as the WLP pathway, consistent with a strictly aerobic mode of metabolism.

## 5.6 Discussion

Multiple notable differences were observed between Zodletone Myxococcota and soil Myxococcota. Of these, the observed variation in GC content was most notable. Within the Bacteria, GC content has been observed to be positively correlated to genome size [8], temperature [269, 417], as well as salinity preferences [301]. Further, higher GC content is associated with aerobic life style[279]. As such, we speculate that the notable difference in GC content between Zodletone Myxococcota and soil Myxococcota could be a reflection of their smaller genome size and aerobic lifestyle. As well, it is possible that a lower rate of intragenic recombination could be occurring in Zodletone Myxococcota leading to a tampered effect of GC-Biased Gene Conversion on genome GC content [225], and a more pronounced effect of spontaneous cytidine deamination on GC to AT mutations [114, 162] collectively leading to a reduction of genomic GC content over evolutionary times from a possible common ancestor with higher GC content as previously speculated [225].

Further, our comparative genomics and machine learning analyses revealed severely curtailed machineries for predation and cellular differentiation in Zodletone Myxococcota (Figures 2, 3, Tables 2 and S6). Such conclusion is in-agreement with a recent study that predicted absence of predation potential in MAGs/SAGs encompassing most of the publicly available, yet-uncultured Myxococcota [390]. As such, a clear delineation exists between two phylogenetically and behaviorally distinct groups within the Myxococcota. The first encompasses aerobic top soil dwellers in classes Myxococcia and Polyangia that are characterized by possessing a highly sophisticated machinery enabling predation and cellular differentiation behaviors. Few freshwater and marine strains possessing such capacities have been reported, but their presence has been attributed to air and dust transport from neighboring soils [131, 132]. Members of this group could readily be obtained in pure cultures. The second group encompasses phylogenetically distinct families and orders within the classes Myxococcia and Polyangia (including Zodletone MAGs), as well a few yet-uncultured Myxococcota classes. These lineages are almost invariably encountered within non-soil habitats (e.g. freshwater, marine, host-associated, and engineered ecosystems), and appear to lack the capacity for predation and social differentiation. Most of these lineages are currently uncultured, with the exception of members of the genus *Anaeromyxobacter* [380].

We argue that such patterns could provide important clues on the evolution of social behavior in the Myxococcota, when considered in light of our understanding of the history

of soil formation and the rise of atmospheric oxygen in the atmosphere. Soil formation and transition from barren crusts to current soil orders through organic matter deposition and transformation has been enabled by the evolution of lichen associations, plant terrestrialization, formation of mycorrhizal association, and subsequent colonization by soil microfauna. All such processes are mediated by aerobic organisms (algae, fungi, plants, fauna), and hence was possible only after the accumulation of oxygen to levels comparable to current values in the atmosphere (approximately 500-600 Mya [55]). The formation of soil structures as a new and organic-rich habitat has certainly spurred multiple evolutionary processes for enabling terrestrial adaptation within the microbial world. Various processes have been reported in multiple soil-prevalent lineages, from CAZymes and BGCs acquisition in the Acidobacteria [112, 407], to acquisition of stress tolerance, adherence, and regulatory genes in the ammonia-oxidizing archaea [2]. Here, it appears that the development of predation and cellular differentiation machineries has enabled the Myxococcota to assume an apex predator niche and imparted them with strong survival capacities in soil, respectively. Indeed, as previously noted, the ecological success of social Myxococcota in soils appears to be in stark contrast to their rarity and low relative abundance of non-social Myxococcota in other habitats [418].

Evolution of beneficial trait(s) in a single lineage of Myxococcota in soil could be propagated to the broader soil Myxococcota community through intra-clade, habitat-specific horizontal gene transfer (HGT), resulting in the observed checkered distribution pattern, where social behavior is observed only in specific families within the classes Myxococcia and Polyangia. HGT between closely related taxa is a well-established phenomenon [3] that has been widely documented, e.g. in mediating the spread of antibiotic resistance in related clinical strains [100, 289]. The barrier for HGT within closely related taxa is predictably lower, given expected similarity in codon usage pattern, GC content, restriction enzyme machinery, and overall genome architecture between donor and recipient strains. Similarly, physical proximity in the same habitat is seen as a facilitator of genetic exchange through HGT [25, 263, 352].

In conclusion, our results strongly indicate that anaerobic Myxococcota do not possess the capacity for typical social behavior, and display distinct structural, anabolic, and catabolic differences when compared to model aerobic Myxococcota. We document their dependence on fermentation and/or nitrite or sulfate-reduction for energy generation, as well as their preference for polysaccharide metabolism over protein, amino acids, and lipid metabolism. We further propose that such differences strongly underscore the importance of niche differentiation in shaping the evolutionary trajectory of the Myxococcota, and suggest soil formation as a strong driver for developing social behavior in this lineage.

## 5.7 Acknowledgements

We thank Dr. David W. Waite at the Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, University of Queensland, for helpful discussions regarding utilization of machine learning approaches for trait prediction. This work was supported by NSF grant 2016423 to NHY and MSE.

This work is published and available in full online (DOI: 10.1128/AEM.01706-21).

## 5.8 Figures & Tables

Supplementary Figures and Tables can be viewed online at:
https://journals.asm.org/doi/10.1128/AEM.01706-21

Table 1. Similarity statistics and Gtdb classification of the MAGs analyzed in this study.

| Bin name | Assembly accession number | % completeness | % contamination | GTDB classification | | | | Similarity statistics | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Class | Order | Family | Genus | RED value | AAI | SGC |
| ZodW_Metabat.174 | JAFGVO01 | 91.24 | 3.26 | Polyangia | JAFGXQ01 | JAFGVO01 | JAFGVO01 | 0.490 | 38.1 ± 0.9 | 40.1 ± 5.62 |
| Zod_Metabat.76 | JAFGQN01 | 94.84 | 3.87 | | | JAFGXQ01 | JAFGQN01 | 0.493 | | |
| ZodW_Metabat.284 | JAFGWT01 | 89.08 | 3.87 | | | | | 0.489 | | |
| Zod_Metabat.947 | JAFGTB01 | 94.19 | 3.23 | | | | JAFGXQ01 | 0.491 | | |
| **ZodW_Metabat.4** | **JAFGXQ01** | **96.01** | **3.23** | | | | | 0.491 | | |
| **Zod_Metabat.169** | **JAFGIB01** | **90.7** | **1.94** | | Polyangiales | JAFGIB01 | JAFGIB01 | 0.660 | 40.1 ± 1.7 | 40.7 ± 5.07 |

RED value: Relative Evolutionary Divergence values, based on placing the genomes in the Gtdb backbone bacterial tree (available at https://data.gtdb.ecogenomic.org/releases/release95/95.0/). RED values were used to confirm the novelty of the taxonomic lineage to which the genomes are assigned. Genomes are assigned to novel orders if their RED values fall between 0.62 and 0.46, and novel families if their RED values fall between 0.62 and 0.77.

AAI: average amino acid identity calculated using the AAI calculator [http://enve-omics.ce.gatech.edu/]). The arbitrary AAI cutoffs used were 49%, 52%, 56%, and 68% for class, order, family, and genus, respectively.

SGC: shared gene content

Table 5.1: Similarity statistics and GTDB classification of the MAGs analyzed in this study

Table 2. The thirteen different pathways/processes examined in this study as determinants of the social behavior in model Myxococcota genomes, along with the number of genes in each module, and the percentage of genes identified in Zodletone genomes as well as the genomes of three Myxococcota type species known to exhibit non-social behavior. More details are in the supplementary text and Table S6.

| **Process/ pathway** | **Major function in the social lifestyle** | **Number of genes in module** | **Zodletone lineages** | | **Type species (non-social)** | | | **Notes on homologues identified/missing in Zodletone genomes** | |
|---|---|---|---|---|---|---|---|---|---|
| | | | **O__JAFGXQ01** | **F__JAFGIB01** | ***Anaeromyxobacter dehalogenes*** | ***Labilithrix luteola*** | ***Vulgatibacter incomptus*** | **Homologues identified** | **Homologues missing** |
| **Enhancer-Binding Proteins (EBP) module** | Gene regulatory networks governing early events of aggregation and mound formation prior to sporulation. | 9 | 88.9 | 88.9 | 88.9 | 100 | 88.9 | • 8 homologues identified including 6 DNA-binding transcriptional response regulator domains of the NtrC family, 1 Serine/Threonine kinases, and 1 peptidase. | • ActC is missing (one of the Act group of proteins) |
| **MrpC module** | | 8 | 75 | 62.5 | 75 | 75 | 87.5 | • 6 homologues identified including catalytic domains of serine/threonine kinases (n=2), two component system sensor histidine kinases (n=2), an oligopeptide transporter (n=1), and a DNA-binding transcriptional response regulator of the NtrC family (n=1). | • MrpC, a transcription factor that works cooperatively to control the start of sporulation, is missing |
| **Nla24 module** | | 2 | 100 | 100 | 100 | 100 | 100 | • 2 homologues identified including a DNA-binding transcriptional response regulator of the NtrC family, and a Diguanylate-cyclase (DGC) or GGDEF domain. | |
| **FruA module** | | 1 | 0 | 0 | 0 | 0 | 0 | | • FruA, a transcription factor that works cooperatively to control the start of sporulation, is missing. |
| **EPS production** | Exopoly-saccharide production necessary for | 8 | 12.5 | 12.5 | 50 | 100 | 62.5 | • 1 homologue identified with a sugar transporter domain. | • 7 out of 8 homologues missing |

Table 5.2: The 13 different pathways/processes examined in this study as determinants of the social behavior in model Myxococcota genomes, along with the number of genes in each module, and the percentage of genes identified in Zodletone genomes as well as the genomes of three Myxococcota type species known to exhibit nonsocial behavior.

| | | | JAFGXQ01 | JAFGIB01 | *Haliangium ochraceum* | *Anaeromyxobacter dehalogenans 2CP-C* | *Melittangium boletus* | *Archangium gephyra* | *Labilithrix luteola* | *Cystobacter fuscus* | *Chondromyces crocatus* | *Persicimonas caui* | *Sorangium cellulosum So ce56* | *Bradymonas sediminis* | *Corallococcus coralloides* | *Sandaracinus amylolyticus* | *Stigmatella aurantiaca* | *Minicystis rosea* | *Myxococcus xanthus* | *Vulgatibacter incomptus* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Carbon sources** | | | | | | | | | | | | | | | | | | | | |
| **Sugar catabolism to central metabolites** | Glucose | | Y | P | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | Mannose | | Y | n | n | Y | Y | Y | n | Y | n | n | Y | n | Y | n | Y | n | Y | Y |
| | Galactose | | Y | n | n | n | n | n | n | n | n | n | n | n | n | Y | n | n | n | n |
| | Fructose | | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | n | Y | Y | Y | Y | Y | n |
| | Hexosamines | | Y | Y | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n |
| | Uronic acids | | n | Y | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n |
| | Mannitol | | Y | n | n | n | n | n | n | Y | n | n | n | n | n | n | n | n | n | n |
| | Arabinose | | Y | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n |
| | Xylose | | Y | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n |
| **Amino acid catabolism** | Ala | | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | Asp | | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | Asn | | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | Glu | | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | Gln | | Y | n | Y | Y | Y | Y | Y | Y | Y | Y | Y | n | Y | Y | Y | Y | Y | Y |
| | Gly | | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | Cys | | Y | n | Y | Y | Y | Y | Y | Y | Y | n | Y | Y | Y | Y | Y | Y | Y | Y |
| | Met | | n | n | Y | n | n | Y | Y | n | n | Y | n | n | n | n | Y | n | n | Y |
| | Ser | | Y | n | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | Thr | | Y | n | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | Val | | n | P | P | P | P | P | P | P | P | P | P | P | P | P | P | P | P | P |
| | Leu | | n | P | P | P | P | P | P | P | P | P | P | P | P | P | P | P | P | P |
| | Ile | | n | Y | P | P | P | P | Y | P | P | Y | P | Y | P | P | P | P | P | P |
| | Lys | | n | n | Y | Y | Y | Y | Y | n | Y | Y | Y | Y | Y | n | Y | Y | Y | Y |
| | Pro | | n | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | Arg | | n | Y | Y | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n |
| | His | | n | n | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | Tyr | | n | n | P | P | P | P | P | P | P | P | P | P | P | P | Y | P | P | P |
| | Trp | | n | n | P | P | n | P | n | n | n | P | n | n | P | n | P | n | n | n |
| **Fatty acids degradation** | Long chain | | P | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | Lactate | | n | n | Y | Y | n | Y | Y | n | n | Y | n | n | n | n | n | n | n | n |
| | Propionate | | Y | Y | n | Y | n | n | Y | n | Y | Y | n | Y | n | n | n | Y | n | n |
| | Butanoate | | n | n | n | Y | n | n | n | n | n | Y | n | n | n | n | n | n | n | n |
| **Fermentation and Respiration** | | | | | | | | | | | | | | | | | | | | |
| **Fate of pyruvate** | Pyruvate to acetyl-CoA | Pyruvate dehydrogenase [EC:1.2.4.1] | n | n | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | | Pyruvate:ferrodoxin oxidoreductase [EC:1.2.7.1] | Y | Y | Y | Y | n | n | n | n | n | n | n | n | n | n | n | n | n | Y |
| | Acetyl-CoA to acetate | Acetyl-CoA synthetase [EC:6.2.1.1] | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | | acetate—CoA ligase (ADP-forming) subunit alpha [EC:6.2.1.13] | Y | Y | Y | n | n | n | Y | n | n | n | n | n | n | n | n | n | n | n |
| | | succinyl-CoA:acetate CoA-transferase [EC:2.8.3.18] | n | n | n | Y | Y | n | n | n | n | n | n | n | n | n | n | n | n | n |
| | | Phosphate acetyltransferase and acetate kinase [EC:2.3.1.8 and EC:2.7.2.1] | n | n | n | Y | n | n | n | n | n | Y | n | n | n | n | n | n | n | n |
| | Ethanol production from acetate | | Y | Y | n | n | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | Ethanol production from acetyl-CoA | | n | n | Y | n | n | n | n | n | n | Y | n | n | n | n | n | Y | n | n |
| | Formate production from pyruvate | | Y | Y | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n |
| | D-lactate production from pyruvate | | Y | Y | Y | n | n | n | n | Y | Y | n | n | n | n | n | n | n | Y | Y |
| | L-lactate production from pyruvate | | n | n | n | n | Y | Y | n | Y | n | n | Y | n | n | Y | Y | Y | n | n |
| | Acetoin production from pyruvate (via acetolactate) | | n | Y | n | Y | n | n | n | n | n | n | Y | n | n | n | n | n | n | n |
| | Butanediol production from acetoin | | n | Y | n | Y | n | n | n | n | n | n | Y | n | n | n | n | n | n | n |
| **TCA cycle** | | | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| **Respiration (aerobic)** | NADH dehydrogenase (complex I) | | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | Succinate dehydrogenase (complex II) | | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | Cytochrome c reductase (complex III) | | n | n | Y | n | n | n | Y | n | n | n | Y | n | n | n | n | n | n | n |
| | Alternate Complex III | | n | n | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | Complex IV | Cytochrome bd respiratory O2 reductase (high O2 affinity, complex IV) | Y | Y | n | Y | n | Y | Y | n | Y | Y | Y | n | Y | n | n | Y | Y | Y |
| | | Cytochrome C oxidase aa3 type (low O2 affinity) | n | n | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | | Cytochrome C oxidase cbb3 type (high O2 affinity) | n | n | n | Y | Y | Y | n | n | Y | n | Y | n | Y | n | Y | Y | Y | Y |
| | ATP synthase (complex V) | F-type | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| **Dissimilatory nitrite reduction to ammonia (DNRA)** | Nitrate reductase | Periplasmic NapAB [EC:1.9.6.1] | n | n | n | n | n | Y | n | n | n | Y | Y | Y | n | n | n | n | n | n |
| | | Membrane-bound NarGHI [EC:1.7.5.1] | n | n | n | Y | n | n | n | n | n | n | n | n | n | n | n | n | n | n |
| | Nitrite reductase | NirBD; NADH [EC:1.7.1.15] | n | n | n | n | n | n | n | Y | n | n | Y | n | n | Y | n | n | n | n |
| | | NrfAH; Cytochrome; ammonia-forming [EC:1.7.2.2] | Y | Y | n | Y | n | Y | n | n | n | n | Y | n | Y | n | n | n | Y | Y |
| **Denitrification** | | | n | n | n | P | n | n | n | n | n | n | n | n | n | n | n | n | n | n |
| **Dissimilatory sulfate reduction** | Sulfate adenylyltransferase | | n | Y | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n |
| | Adenylylsulfate reductase | | n | Y | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n |
| | Quionone interacting membrane bound oxidoreductase complex (QmoABC) | | n | Y | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n |
| | Dissimilatory sulfite reductase | | n | Y | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n |
| **Tetrathionate respirattion** | Octaheme tetrathionate reductase (Otr) | | n | Y | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n | n |
| **Thiosulfate disproportionation** | Thiosulfate reductase (PhsABC) | | n | Y | n | Y | n | n | n | n | n | n | n | n | n | n | n | n | n | n |
| **Thiosulfate ox** | SOX complex | | n | n | n | Y | n | n | n | n | n | n | n | n | n | n | n | n | n | n |
| **Sulfite oxidation** | SoeABC | | n | n | n | Y | n | n | n | n | n | n | n | n | n | n | n | n | n | n |
| **Sulfide oxidation** | sulfide:quinone oxidoreductase [EC:1.8.5.4] | | n | n | Y | Y | Y | Y | n | Y | Y | Y | n | n | Y | n | Y | n | Y | n |

Table 5.3: Catabolic capabilities deduced from genomic analysis of Zodletone Myxococcota in comparison to type species genomes available from the KEGG organism database

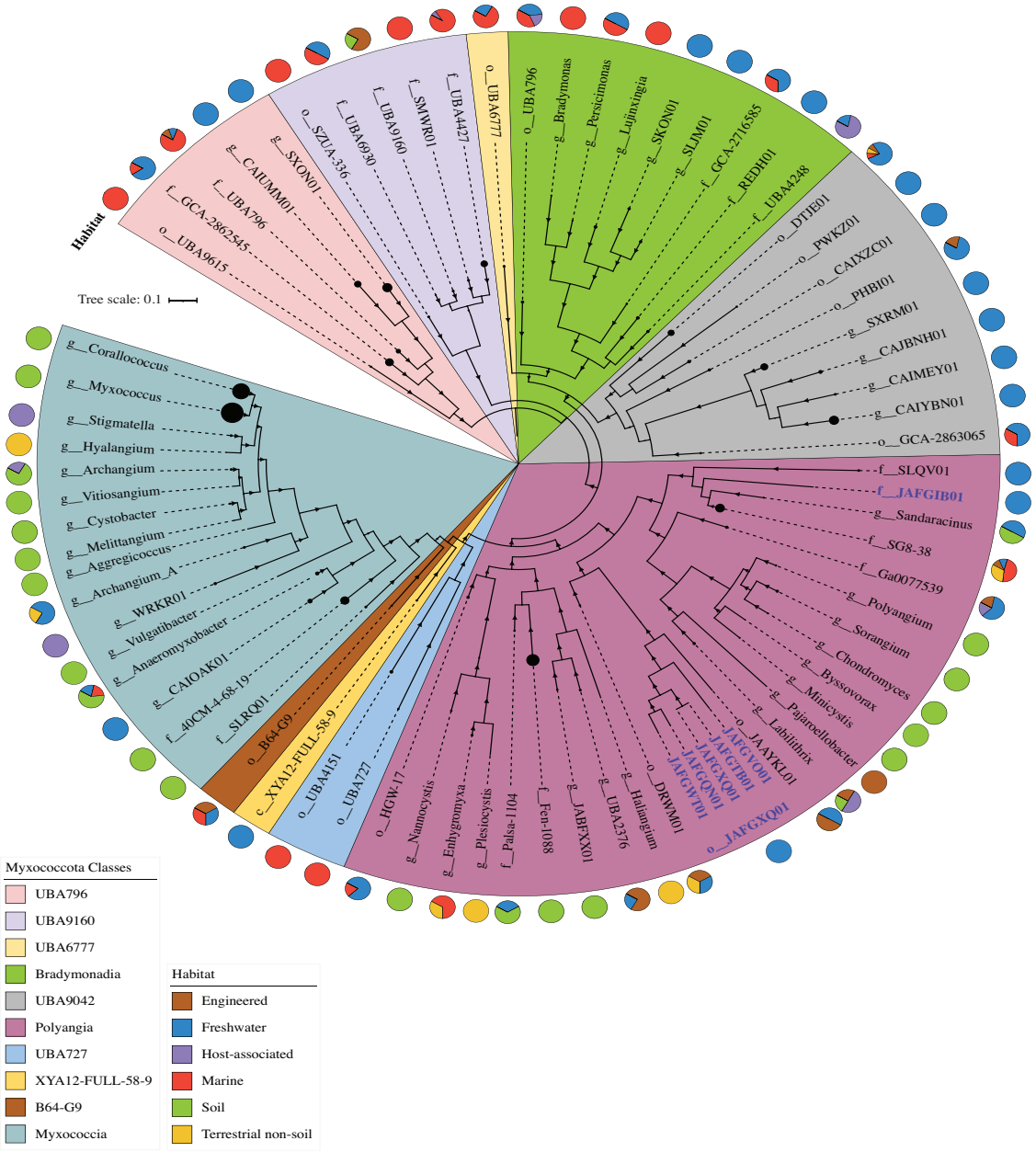



Figure 5.1: Phylogenomics of the Myxococcota, including novel lineages from Zodletone Spring. The maximum likelihood trees were constructed in RAxML [369] using all *Myxococcota* genomes available from the GTDB r95 database based on the concatenated alignments of 120 housekeeping genes obtained from GTDB-Tk [65]. The tree was rooted (root not shown) with the two *Bdellovibrionota* genomes *Halobacteriovorax marinus* (GenBank assembly accession number GCF_000210915.2) and *Bdellovibrio bacteriovorus* (GenBank assembly accession number GCF_000196175.1). The tree is wedged (shown as black circles at the end of branches) to represent genus level taxonomy (g__), unless the number of available genomes per genus is less than 5, in which case the family level (f__), or order level (o__) taxonomy is shown instead. The size of the wedge is proportional to the number of genomes. Bootstrap support values based on 100 replicates are shown as triangles for nodes with ¿70% support. Class-level taxonomy is color-coded as shown in the legend. The track around the tree represents the ecosystem classification of the habitat from which the genomes originated. Zodletone genomes are labeled in blue bold text with their GenBank assembly accession number.

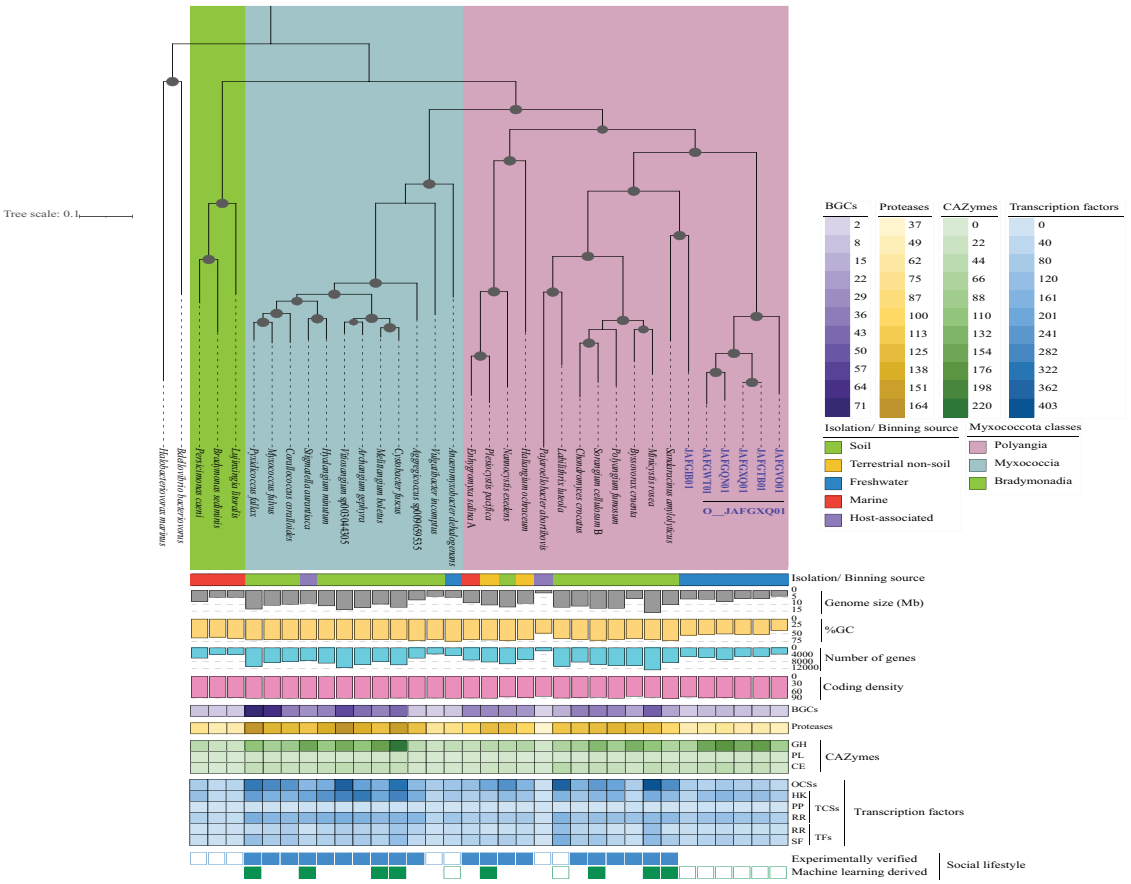



Figure 5.2: Comparative genomics of Zodletone novel *Myxococcota* genomes to the genomes of 27 type species belonging to the classes *Myxococcia*, *Polyangia*, and *Bradymonadia*. The species and their GenBank assembly accession numbers are *Anaeromyxobacter dehalogenans* 2CP-1, GCF_000022145.1; *Haliangium ochraceum* DSM 14365, GCF_000024805.1; *Plesiocystis pacifica* SIR-1, GCF_000170895.1; *Corallococcus coralloides* DSM 2259, GCF_000255295.1; *Cystobacter fuscus* DSM 2262, GCF_000335475.2; *Hyalangium minutum*, GCF_000737315.1; *Sandaracinus amylolyticus*, GCF_000737325.1; *Archangium gephyra*, GCF_001027285.1; *Chondromyces crocatus*, GCF_001189295.1; *Vulgatibacter incomptus*, GCF_001263175.1; *Labilithrix luteola*, GCF_001263205.1; *Minicystis rosea*, GCF_001931535.1; *Melittangium boletus* DSM 14713, GCF_002305855.1; *Nannocystis exedens*, GCF_002343915.1; *Bradymonas sediminis*, GCF_003258315.1; *Lujinxingia litoralis*, GCF_003260125.1; *Polyangium fumosum*, GCF_005144585.1; *Persicimonas caeni*, GCF_006517175.1; *Myxococcus fulvus*, GCF_007991095.1; *Pyxidicoccus fallax*, GCF_012933655.1; *Stigmatella aurantiaca*, GCF_900109545.1; i, GCF_003044305.1; *Aggregicoccus*, GCA_009659535.1; *Pajaroellobacter abortibovis*, GCF_001931505.1; *Byssovorax cruenta*, GCA_001312805.1; *Enhygromyxa salina*, GCF_002994615.1; and *Sorangium cellulosum* B, GCF_000067165.1. Zodletone genomes are labeled in blue bold text with their GenBank assembly accession number. Class-level taxonomy is color-coded as shown in the legend. The tracks underneath the tree show the ecosystem classification of the habitat from which the genomes originated, the assembly genome size (gray bars), GC content (yellow bars), total number of genes in the genome (cyan bars), and coding density (pink bars). The number of biosynthetic gene clusters (BGCs, purple), proteases (yellow), CAZymes (green), and transcription factors (blue) encoded in each genome are shown as a heatmap with the color tones explained in the legend. Under the heatmaps, the two outermost tracks denote the presence (filled squares)/absence (empty squares) of the *Myxococcota* typical social lifestyle as evidenced by experimental pure culture work (blue), and/or the machine learning approach (green) we used for lifestyle prediction based on the informative set of KO numbers provided in Data Set S1. BGCs, biosynthetic gene clusters; GH, glycosyl hydrolases; PL, polysaccharide lyases; CE, carbohydrate esterase; OCSs, one-component systems; TFs, transcription factors; RR, response regulator; SF, sigma factor; TCSs, two-component systems; HK, histidine kinases; PP, phospho-relay proteins.

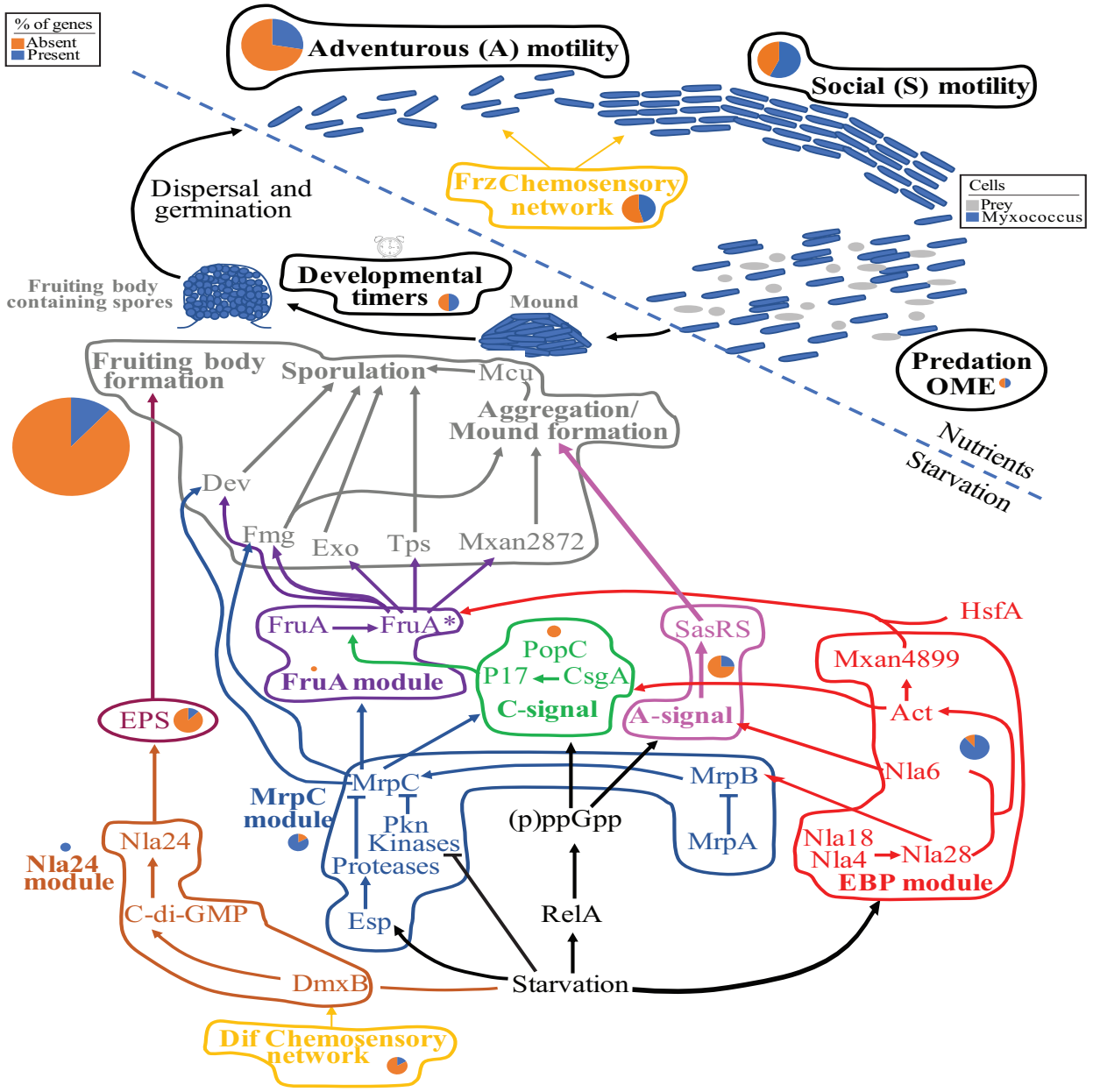



Figure 5.3: A cartoon depicting the 13 pathways associated with *Myxococcota* social lifestyle examined in detail in this study. *Myxococcus* cells are shown as blue rods, while prey cells are depicted as gray cocci and rods. Pathways active during nutrient availability are shown above the dotted line, while those induced by starvation are shown below the dotted line. Each of the pathways is shown in bold text within a color-coded outline. The same color code is used for the group of genes in each pathway and in Table S6. For each pathway, a pie chart for the number of gene homologues identified in Zodletone genomes as a percentage of the total number of genes in the pathway is shown in blue, while the percentage of gene homologues absent is shown in orange. The size of the pie chart is proportional to the number of genes in each pathway and ranges from 1 (FruA module) to 35 (the aggregation/sporulation/fruiting body formation module). Arrowheads depict the effect where activation is shown as triangular arrowheads, and inhibition is shown as horizontal line arrowheads. FruA* denotes the active form of FruA. EPS, exopolysaccharide; OME, outer membrane exchange.

## REFERENCES


[1] L. K. Abbott and N. C. Johnson, *Chapter 6 - introduction: Perspectives on mycorrhizas and soil fertility*, pp. 93–105, Elsevier, 2017.

[2] Sophie S. Abby, Melina Kerou, Christa Schleper, Derek R. Lovley, Kira Makarova, and Patrick Forterre, *Ancestral reconstructions decipher major adaptations of ammonia-oxidizing archaea upon radiation into moderate terrestrial and marine environments*, mBio **11** (2020), no. 5, e02371–20.

[3] Orit Adato, Noga Ninyo, Uri Gophna, and Sagi Snir, *Detecting horizontal gene transfer between closely related taxa*, PLOS Computational Biology **11** (2015), no. 10, e1004408.

[4] W. G. Alexander, J. H. Wisecaver, A. Rokas, and C. T. Httinger, *Horizontally aquired gene in early-diverging pathogenic fungi enable the use of host nucleosides and nucleotides*, Proc. Nat. Acad. Sci. USA **113** (2016), 4116–4121.

[5] J. P. Allen and J. C. Williams, *Reaction centers from purple bacteria*, pp. 275–293, Wiley-Blackwell, Weinheim, Germany, 2008.

[6] Jeffrey R. Allen, Daniel D. Clark, Jonathan G. Krum, and Scott A. Ensign, *A role for coenzyme m (2-mercaptoethanesulfonic acid) in a bacterial pathway of aliphatic epoxide carboxylation*, Proceedings of the National Academy of Sciences **96** (1999), no. 15, 8432–8437.

[7] Alexandre Almeida, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S. Pollard, Ekaterina Sakharova, Donovan H. Parks, Philip Hugenholtz, Nicola Segata, Nikos C. Kyrpides, and Robert D. Finn, *A unified catalog of 204,938 reference genomes from the human gut microbiome*, Nature Biotechnology **39** (2021), no. 1, 105–114.

[8] A. Almpanis, M. Swain, D. Gatherer, and N. McEwan, *Correlation between bacterial g+c content, genome size and the g+c content of associated plasmids and bacteriophages*, Microb Genom **4** (2018), no. 4, NULL.

[9] K. Anantharaman, B. Hausmann, S. P. Jungbluth, R. S. Kantor, A. Lavy, L. A. Warren, M. S. Rappe, M. Pester, A. Loy, B. C. Thomas, and J. F. Banfield, *Expanded diversity of microbial groups that shape the dissimilatory sulfur cycle*, The ISME J. **12** (2018), 1715–1728.

[10] Karthik Anantharaman, Christopher T. Brown, Laura A. Hug, Itai Sharon, Cindy J. Castelle, Alexander J. Probst, Brian C. Thomas, Andrea Singh, Michael J. Wilkins, Ulas Karaoz, Eoin L. Brodie, Kenneth H. Williams, Susan S. Hubbard, and Jillian F. Banfield, *Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system*, Nature Communications **7** (2016), no. 1, 13219.

[11] D. I. Andersson, J. Jerlstrom-Hultqvist, and J. Nasvall, *Evolution of new functions de novo and from preexisting genes*, Cold Spring Harb Perspect Biol **7** (2015), no. 6, NULL.

[12] D. I. Andersson, J. Jerlström-Hultqvist, and J. Näsvall, *Evolution of new functions de novo and from preexisting genes*, Cold Spring Harb Perspect Biol. **7** (2015), a017996.

[13] Jordan C. Angle, Timothy H. Morin, Lindsey M. Solden, Adrienne B. Narrowe, Garrett J. Smith, Mikayla A. Borton, Camilo Rey-Sanchez, Rebecca A. Daly, Golnazalsdat Mirfenderesgi, David W. Hoyt, William J. Riley, Christopher S. Miller, Gil Bohrer, and Kelly C. Wrighton, *Methanogenesis in oxygenated soils is a substantial fraction of wetland methane emissions*, Nature Communications **8** (2017), no. 1, 1567.

[14] S. M. Arellano, O. O. Lee, F. F. Lafi, J. Yang, Y. Wang, C. M. Young, and P. Y. Qian, *Deep sequencing of myxilla (ectyomyxilla) methanophila, an epibiotic sponge on cold-seep tubeworms, reveals methylotrophic, thiotrophic, and putative hydrocarbon-degrading microbial associations*, Microb Ecol **65** (2013), no. 2, 450–61.

[15] A. J. Auman and M. E. Lidstrom, *Analysis of smmo-containing type i methanotrophs in lake washington sediment*, Environ Microbiol **4** (2002), no. 9, 517–24.

[16] A. J. Auman, S. Stolyar, A. M. Costello, and M. E. Lidstrom, *Molecular characterization of methanotrophic isolates from freshwater lake sediment*, Appl Environ Microbiol **66** (2000), no. 12, 5259–66.

[17] Luke D. Bainard, Jillian D. Bainard, Chantal Hamel, and Yantai Gan, *Spatial and temporal structuring of arbuscular mycorrhizal communities is differentially influenced by abiotic factors and host crop in a semi-arid prairie agroecosystem*, FEMS Microbiology Ecology **88** (2014), no. 2, 333–344.

[18] Luke D. Bainard, Alexander M. Koch, Andrew M. Gordon, Steven G. Newmaster, Naresh V. Thevathasan, and John N. Klironomos, *Influence of trees on the spatial structure of arbuscular mycorrhizal communities in a temperate tree-based intercropping system*, Agriculture, Ecosystems & Environment **144** (2011), no. 1, 13–20.

[19] R. Balasubramanian, S. M. Smith, S. Rawat, L. A. Yatsunyk, T. L. Stemmler, and A. C. Rosenzweig, *Oxidation of methane by a biological dicopper centre*, Nature **465** (2010), no. 7294, 115–9.

[20] W. E. Balch and R. Wolfe, *New approach to the cultivation of methanogenic bacteria: 2-mercaptoethanesulfonic acid (hs-com)-dependent growth of methanobacterium ruminantium in a pressureized atmosphere*, Appl. Environ. Microbiol. **32** (1976), 781–791.

[21] M. S. Bansal, M. Kellis, M. Kordi, and S. Kundu, *Ranger-dtl 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss*, Bioinformatics **34** (2018), no. 18, 3214–3216.

[22] S. C. Bayliss, H. A. Thorpe, N. M. Coyle, S. K. Sheppard, and E. J. Feil, *Pirate: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria*, Gigascience **8** (2019), no. 10, NULL.

[23] J. P. Beam, E. D. Becraft, K. M. Brown, F. Schulz, and J. K. Jarett, *Ancestral absence of electron transport chains in patescibacteria and dpann*, Front. Micorobiol. (2020), https://doi.org/10.3389/fmicb.2020.01848.

[24] E. D. Becraft, T. Woyke, J. Jarett, N. Ivanova, F. Godoy-Vitorino, N. Poulton, J. M. Brown, J. Brown, M. C. Y. Lau, T. Onstott, J. A. Eisen, D. Moser, and R. Stepanauskas, *Rokubacteria: Genomic giants among theuncultured bacterial phyla.*, Front. Micorobiol. **8** (2017), 2264.

[25] R. G. Beiko, T. J. Harlow, and M. A. Ragan, *Highways of gene sharing in prokaryotes*, Proc Natl Acad Sci U S A **102** (2005), no. 40, 14332–7.

[26] Mary L. Berbee, Timothy Y. James, and Christine Strullu-Derrien, *Early diverging fungi: Diversity and impact at the dawn of terrestrial life*, Annual Review of Microbiology **71** (2017), no. 1, 41–60.

[27] J. E. Berleman and J. R. Kirby, *Deciphering the hunting strategy of a bacterial wolfpack*, FEMS Microbiol Rev **33** (2009), no. 5, 942–57.

[28] S. Bikel, A. Valdez-Lara, and F. Cornejo-Granados, *Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome.*, Comput Struct Biotechnol J. **13** (2015), 390–401.

[29] P. J. Bishop, R. Speare, R. Poulter, M. Butler, B. J. Speare, A. Hyatt, V. Olsen, and A. Haigh, *Elimination of the amphibian chytrid fungus batrachochytrium dendrobatidis by archey's frog leiopelma archeyi*, Dis Aquat Organ **84** (2009), no. 1, 9–15.

[30] Michael Bitterlich, Martin Sandmann, and Jan Graefe, *Arbuscular mycorrhiza alleviates restrictions to substrate water flow and delays transpiration limitation to stronger drought in tomato*, Frontiers in plant science **9** (2018), 154–154.

[31] B. Blagoev, S. E. Ong, I. Kratchmarova, and M. Mann, *Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics*, Nat Biotechnol **22** (2004), no. 9, 1139–45.

[32] R. Boden, E. Thomas, P. Savani, D. P. Kelly, and A. P. Wood, *Novel methylotrophic bacteria isolated from the river thames (london, uk)*, Environ Microbiol **10** (2008), no. 12, 3225–36.

[33] A. M. Bolger, M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for illumina sequence data*, Bioinformatics **30** (2014), no. 15, 2114–20.

[34] T. C. Boothby, J. R. Tenlen, F. W. Smith, J. R. Wang, K. A. Patanella, E. O. Nishimura, S. C. Tintori, Q. Li, C. D. Jones, M. Yandell, D. N. Messina, J. Glasscock, and B. Goldstein, *Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade*, Proc. Nat. Acad. Sci. USA **112** (2015), 15976–15981.

[35] A. V. Borges, G. Abril, B. Delille, J.-P. Descy, and F. Darchambeau, *Diffusive methane emissions to the atmosphere from lake kivu (eastern africa)*, Journal of Geophysical Research: Biogeosciences **116** (2011), no. G3, NULL.

[36] C. Boschetti, A. Carr, A. Crisp, I. Eyres, Y. Wang-Koh, E. Lubzens, T. G. Barraclough, G. Micklem, and A. Tunnacliffe, *Biochemical diversification through foreign gene expression in bdelloid rotifers*, PLOS Genet. **8** (2012), e1003035.

[37] R. M. Bowers, N. C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T. B. K. Reddy, F. Schulz, J. Jarett, A. R. Rivers, E. A. Eloe-Fadrosh, S. G. Tringe, N. N. Ivanova, A. Copeland, A. Clum, E. D. Becraft, R. R. Malmstrom, B. Birren, M. Podar, P. Bork, G. M. Weinstock, G. M. Garrity, J. A. Dodsworth, S. Yooseph, G. Sutton, F. O. Glöckner, J. A. Gilbert, W. C. Nelson, S. J. Hallam, S. P. Jungbluth, T. J. G. Ettema, S. Tighe, K. T. Konstantinidis, W. T. Liu, B. J. Baker, T. Rattei, J. A. Eisen, B. Hedlund, K. D. McMahon, N. Fierer, R. Knight, R. Finn, G. Cochrane, I. Karsch-Mizrachi, G. W. Tyson, C. Rinke, A. Lapidus, F. Meyer, P. Yilmaz, D. H. Parks, A. M. Eren, L. Schriml, J. F. Banfield, P. Hugenholtz, and T. Woyke, *Minimum information about a single amplified genome (misag) and a metagenome-assembled genome (mimag) of bacteria and archaea*, Nat Biotechnol **35** (2017), no. 8, 725–731.

[38] T. M. Bowles, L. E. Jackson, and T. R. Cavagnaro, *Mycorrhizal fungi enhance plant nutrient acquisition and modulate nitrogen loss with variable water regimes*, Glob Chang Biol **24** (2018), no. 1, e171–e182.

[39] JOHN P. BOWMAN, LINDSAY I. SLY, PETER D. NICHOLS, and A. C. HAYWARD, *Revised taxonomy of the methanotrophs: Description of methylobacter gen. nov., emendation of methylococcus, validation of methylosinus and methylocystis species, and a proposal that the family methylococcaceae includes only the group i methanotrophs*, International Journal of Systematic and Evolutionary Microbiology **43** (1993), no. 4, 735–753.

[40] B. Boxma, F. Voncken, S. Jannink, T. Van Alen, A. Akhmanova, S. W. H. Van Weelden, J. J. Van Hellemond, G. Ricard, M. Huynen, A. G. M. Tielens, and J. H. P. Hackstein, *The anaerobic chytridiomycete fungus piromyces sp. e2 produces ethanol via pyruvate:formate lyase and an alcohol dehydrogenase*, E. Mol. Microbiol. **51** (2004), 1389 –1399.

[41] A. Boysen, E. Ellehauge, B. Julien, and L. Søgaard-Andersen, *The devt protein stimulates synthesis of frua, a signal transduction protein required for fruiting body morphogenesis in myxococcus xanthus*, J Bacteriol **184** (2002), no. 6, 1540–6.

[42] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, *Near-optimal probabilistic rna-seq quantification*, Nat. Biotechnol. **34** (2016), 525–527.

[43] A. P. Bretscher and D. Kaiser, *Nutrition of myxococcus xanthus, a fruiting myxobacterium*, J Bacteriol **133** (1978), no. 2, 763–8.

[44] C. T. Brown, M. R. Olm, B. C. Thomas, and J. F. Banfield, *Measurement of bacterial replication rates in microbial communities*, Nat Biotechnol. **34** (2016), 1256–1263.

[45] Mark C. Brundrett, *Mycorrhizal associations and other means of nutrition of vascular plants: understanding the global diversity of host plants by resolving conflicting information and developing reliable means of diagnosis*, Plant and Soil **320** (2009), no. 1, 37–77.

[46] M. Bryant, *Commentary on the hungate technique for culture of anaerobic bacteria.*, Am. J. Clin. Nutr. **25** (1972), 1324–1328.

[47] S. I. Buhring, S. M. Sievert, H. M. Jonkers, T. Ertefai, M. S. Elshahed, L. R. Krumholz, and K. U Hinrichs, *Insights into chemotaxonomic composition and carbon cycling of phototrophic communities in an artesian sulfur-rich spring (zodletone, oklahoma, usa), a possible analog for ancient microbial mat systems*, Geobiology **9** (2011), 166–179.

[48] C. N. Butterfield, Z. Li, P. F. Andeer, S. Spaulding, B. C. Thomas, A. Singh, R. L. Hettich, K. B. Suttle, A. J. Probst, S. G. Tringe, T. Northen, C. Pan, and J. F. Banfield, *Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone*, PeerJ **4** (2016), e2687.

[49] Guillaume Bécard and Yves Piché, *6 establishment of vesicular-arbuscular mycorrhiza in root organ culture: Review and proposed methodology*, vol. 24, pp. 89–108, Academic Press, 1992.

[50] N. B. Caberoy, R. D. Welch, J. S. Jakobsen, S. C. Slater, and A. G. Garza, *Global mutational analysis of ntrc-like activators in myxococcus xanthus: identifying activator mutants defective for motility and fruiting body development*, J Bacteriol **185** (2003), no. 20, 6083–94.

[51] J. Cai, R. Zhao, H. Jiang, and W. Wang, *De novo origination of a new protein-coding gene in saccharomyces cerevisiae*, Genetics **179** (2008), 487–496.

[52] S. Calkins, N. C. Elledge, R. A. Hanafy, M. S. Elshahed, and N. H. Youssef, *A fast and reliable procedure for spore collection from anaerobic fungi: Application for rna uptake and long-term storage of isolates*, J. Microbiol. Methods. **127** (2016), 206–213.

[53] T. M. Callaghan, S. M. Podmirseg, D. Hohlweck, J. E. Edwards, A. K. Puniya, S. S. Dagar, and G. W. Griffith, *Buwchfawromyces eastonii gen. nov., sp. nov.: a new anaerobic fungus (neocallimastigomycota) isolated from buffalo faeces*, Mycokeys **9** (2015), 11–28.

[54] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, *Blast+: architecture and applications*, BMC Bioinformatics **10** (2009), 421.

[55] D. E. Canfield, *A new model for proterozoic ocean chemistry*, Nature **396** (1998), no. 6710, 450–453.

[56] D. E. Canfield, K. S. Habicht, and B. Thamdrup, *The archean sulfur cycle and the early history of atmospheric oxygen*, Science **299** (2000), 658–661.

[57] L. Cao, O. Caldararu, A. C. Rosenzweig, and U. Ryde, *Quantum refinement does not support dinuclear copper sites in crystal structures of particulate methane monooxygenase*, Angew Chem Int Ed Engl **57** (2018), no. 1, 162–166.

[58] P. Cao, A. Dey, C. N. Vassallo, and D. Wall, *How myxobacteria cooperate*, J Mol Biol **427** (2015), no. 23, 3709–21.

[59] S. Cao, M. Wu, S. Xu, X. Yan, and X. Mao, *Identification of a putative flavin adenine dinucleotide-binding monooxygenase as a regulator for myxococcus xanthus development*, J Bacteriol **197** (2015), no. 7, 1185–96.

[60] Irene M. Cardoso and Thomas W. Kuyper, *Mycorrhizas and tropical soil fertility*, Agriculture, Ecosystems & Environment **116** (2006), no. 1, 72–84.

[61] C. R. Carere, K. Hards, K. M. Houghton, J. F. Power, B. McDonald, C. Collet, D. J. Gapes, R. Sparling, E. S. Boyd, G. M. Cook, C. Greening, and M. B. Stott, *Mixotrophy drives niche expansion of verrucomicrobial methanotrophs*, Isme j **11** (2017), no. 11, 2599–2610.

[62] A. Caro-Quintero and K. Konstantinidis, *Inter-phylum hgt has shaped the metabolism of many mesophilic and anaerobic bacteria*, ISME J **9** (2015), 958–967.

[63] A. R. Carvunis, T. Rolland, I. Wapinski, M. A. Calderwood, M. A. Yildirim, N. Simonis, B. Charloteaux, C. A. Hidalgo, J. Barbette, B. Santhanam, G. A. Brar, J. S. Weissman, A. Regev, N. Thierry-Mieg, M. E. Cusick, and M. Vidal, *Proto-genes and de novo gene birth*, Nature **487** (2010), 370–374.

[64] P. P. Chan and T. M. Lowe, *trnascan-se: Searching for trna genes in genomic sequences*, Methods Mol Biol **1962** (2019), 1–14.

[65] P. A. Chaumeil, A. J. Mussig, P. Hugenholtz, and D. H. Parks, *Gtdb-tk: a toolkit to classify genomes with the genome taxonomy database*, Bioinformatics (2019), NULL.

[66] E. C. Chen, S. Mathieu, A. Hoffrichter, K. Sedzielewska-Toro, M. Peart, A. Pelin, S. Ndikumana, J. Ropars, S. Dreissig, J. Fuchs, A. Brachmann, and N. Corradi, *Single nucleus sequencing reveals evidence of inter-nucleus recombination in arbuscular mycorrhizal fungi*, Elife **7** (2018), NULL.

[67] E. C. H. Chen, E. Morin, D. Beaudet, J. Noel, G. Yildirir, S. Ndikumana, P. Charron, C. St-Onge, J. Giorgi, M. Kruger, T. Marton, J. Ropars, I. V. Grigoriev, M. Hainaut, B. Henrissat, C. Roux, F. Martin, and N. Corradi, *High intraspecific genome diversity in the model arbuscular mycorrhizal symbiont rhizophagus irregularis*, New Phytol **220** (2018), no. 4, 1161–1171.

[68] I-M A Chen, K. Chu, K. Palaniappan, M. Pillay, A. Ratner, J. Huang, M. Huntemann, N. Varghese, J. R White, R. Seshadri, T. Smirnova, E. Kirton, S. P Jungbluth, T. Woyke, E. A. Eloe-Fadrosh, N. N Ivanova, and N. C. Kyrpides, *Img/m v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes*, Nucl. Acids Res. **47** (2019), 666–677.

[69] Q. Chen, D. B. Janssen, and B. Witholt, *Growth on octane alters the membrane lipid fatty acids of pseudomonas oleovorans due to the induction of alkb and synthesis of octanol*, J Bacteriol **177** (1995), no. 23, 6894–901.

[70] Davide Chicco and Giuseppe Jurman, *The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation*, BMC Genomics **21** (2020), no. 1, 6.

[71] L. Chistoserdova, *Methylotrophy in a lake: from metagenomics to single-organism physiology*, Appl Environ Microbiol **77** (2011), no. 14, 4705–11.

[72] _____, *Modularity of methylotrophy, revisited*, Environ Microbiol **13** (2011), no. 10, 2603–22.

[73] _____, *Methylotrophs in natural habitats: current insights through metagenomics*, Appl Microbiol Biotechnol **99** (2015), no. 14, 5763–79.

[74] L. Chistoserdova, M. G. Kalyuzhnaya, and M. E. Lidstrom, *The expanding world of methylotrophic metabolism*, Annu Rev Microbiol **63** (2009), 477–99.

[75] Ludmila Chistoserdova, *The distribution and evolution of c1 transfer enzymes and evolution of the planctomycetes*, pp. 195–209, Humana Press, Totowa, NJ, 2013.

[76] Ludmila Chistoserdova and Mary E. Lidstrom, *Aerobic methylotrophic prokaryotes*, pp. 267–285, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[77] M. Chuvochina, C. Rinke, D. H. Parks, M. S. Rapp, G. W. Tyson, P. Yilmaz, W. B. Whitman, and P. Hugenholtz, *The importance of designating type material for uncultured taxa*, Syst. Appl. Microbiol. **In Press, DOI: https://doi.org/10.1016/j.syapm.2018.07.003** (2018), NULL.

[78] Nicholas V. Coleman and Jim C. Spain, *Epoxyalkane:coenzyme m transferase in the ethene and vinyl chloride biodegradation pathways of ¡em¿mycobacterium¡/em¿ strain js60*, Journal of Bacteriology **185** (2003), no. 18, 5536–5545.

[79] Daniel R. Colman, Melody R. Lindsay, Maximiliano J. Amenabar, Maria C. Fernandes-Martins, Eric R. Roden, and Eric S. Boyd, *Phylogenomic analysis of novel diaforarchaea is consistent with sulfite but not sulfate reduction in volcanic environments on early earth*, The ISME Journal **14** (2020), no. 5, 1316–1331.

[80] R. Conrad, *The global methane cycle: recent advances in understanding the microbial processes involved*, Environ Microbiol Rep **1** (2009), no. 5, 285–92.

[81] Luis M Corrochano, Alan Kuo, Marina Marcet-Houben, Silvia Polaino, Asaf Salamov, José M Villalobos-Escobedo, Jane Grimwood, M. Isabel Álvarez, Javier Avalos, Diane Bauer, Ernesto P Benito, Isabelle Benoit, Gertraud Burger, Lola P Camino, David Cánovas, Enrique Cerdá-Olmedo, Jan-Fang Cheng, Angel Domínguez, Marek Eliáš, Arturo P Eslava, Fabian Glaser, Gabriel Gutiérrez, Joseph Heitman, Bernard Henrissat, Enrique A Iturriaga, B. Franz Lang, José L Lavín, Soo Chan Lee, Wenjun Li, Erika Lindquist, Sergio López-García, Eva M Luque, Ana T Marcos, Joel Martin, Kevin McCluskey, Humberto R Medina, Alejandro Miralles-Durán, Atsushi Miyazaki, Elisa Muñoz-Torres, José A Oguiza, Robin A Ohm, María Olmedo, Margarita Orejas, Lucila Ortiz-Castellanos, Antonio G Pisabarro, Julio Rodríguez-Romero, José Ruiz-Herrera, Rosa Ruiz-Vázquez, Catalina Sanz, Wendy Schackwitz, Mahdi Shahriari, Ekaterina Shelest, Fátima Silva-Franco, Darren Soanes, Khajamohiddin Syed, Víctor G Tagua, Nicholas J Talbot, Michael R Thon, Hope Tice, Ronald P de Vries, Ad Wiebenga, Jagjit S Yadav, Edward L Braun, Scott E Baker, Victoriano Garre, Jeremy Schmutz, Benjamin A Horwitz, Santiago Torres-Martínez, Alexander Idnurm, Alfredo Herrera-Estrella, Toni Gabaldón, and Igor V Grigoriev, *Expansion of signal transduction pathways in fungi by extensive genome duplication*, Current Biology **26** (2016), no. 12, 1577–1584.

[82] M. B. Couger, N. H. Youssef, C. G. Struchtemeyer, A. S. Liggenstoffer, and M. S. Elshahed, *Transcriptomic analysis of lignocellulosic biomass degradation by the anaerobic fungal isolate orpinomyces sp. strain c1a*, Biotechnol Biofuels **8** (2015), 208.

[83] David Couvin, Aude Bernheim, Claire Toffano-Nioche, Marie Touchon, Juraj Michalik, Bertrand Néron, Eduardo P C Rocha, Gilles Vergnaud, Daniel Gautheret, and Christine Pourcel, *Crisprcasfinder, an update of crisrfinder, includes a portable version, enhanced performance and integrates search for cas proteins*, Nucleic Acids Research **46** (2018), no. W1, W246–W251.

[84] J. Cox and M. Mann, *Maxquant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification*, Nat Biotechnol **26** (2008), no. 12, 1367–72.

[85] Sylvie Cranenbrouck, Liesbeth Voets, Céline Bivort, Laurent Renard, Désiré-Georges Strullu, and Stéphane Declerck, *Methodologies for in vitro cultivation of arbuscular mycorrhizal fungi with root organs*, pp. 341–375, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[86] Christopher J. Creevey, William J. Kelly, Gemma Henderson, and Sinead C. Leahy, *Determining the culturability of the rumen bacterial microbiome*, Microbial Biotechnology **7** (2014), no. 5, 467–479.

[87] Alastair Crisp, Chiara Boschetti, Malcolm Perry, Alan Tunnacliffe, and Gos Micklem, *Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes*, Genome Biology **16** (2015), no. 1, 50.

[88] M. T. Croft, A. D. Lawrence, E. Raux-Deery, M. J. Warren, and A. G. Smith, *Algae acquire vitamin b12 through a symbiotic relationship with bacteria*, Nature **438** (2005), no. 7064, 90–3.

[89] Patrick D. Curtis and Lawrence J. Shimkets, *Metabolic pathways relevant to predation, signaling, and development*, pp. 241–258, Wiley Online Library, 2007.

[90] S. S. Dagar, S. Kumar, G. W. Griffith, J. E. Edwards, T. M. Callaghan, R. Singh, A. K. Nagpal, and A. K. Puniya, *A new anaerobic fungus (oontomyces anksri gen. nov., sp. nov.) from the digestive tract of the indian camel (camelus dromedarius)*, Fungal Biol. **19** (2015), 731–737.

[91] M. Dai, L. D. Bainard, C. Hamel, Y. Gan, and D. Lynch, *Impact of land use on arbuscular mycorrhizal fungal communities in rural canada*, Appl Environ Microbiol **79** (2013), no. 21, 6719–29.

[92] E. G. Danchin, M. N. Rosso, P. Vieira, J. de Almeida-Engler, P. M. Coutinho, B. Henrissat, and P. Abad, *Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes*, Proc. Nat. Acad. Sci. USA **107** (2010), 17651–17656.

[93] M. De Cáceres and P. Legendre, *Associations between species and groups of sites: indices and statistical inference*, Ecology **90** (2009), no. 12, 3566–74.

[94] A. P. de Koning, F. S. Brinkman, S. J. Jones, and P. J. Keeling, *Lateral gene transfer and metabolic adaptation in the human parasite trichomonas vaginalis.*, Mol. Biol. Evol. **17** (2000), 1769–1773.

[95] Stéphane Declerck, Désiré G. Strullu, and Christian Plenchette, *Monoxenic culture of the intraradical forms of glomus sp. isolated from a tropical ecosystem: A proposed methodology for germplasm collection*, Mycologia **90** (1998), no. 4, 579–585.

[96] X. Deng, N. Dohmae, K. H. Nealson, K. Hashimoto, and A. Okamoto, *Multi-heme cytochromes provide a pathway for survival in energy-limited environments*, Sci Adv **4** (2018), no. 2, eaao5682.

[97] S. Diamond, P. F. Andeer, Z. Li, A. Crits-Christoph, D. Burstein, K. Anantharaman, K. R. Lane, B. C. Thomas, C. Pan, T. R. Northen, and J. F. Banfield, *Mediterranean grassland soil c-n compound turnover is dependent on rainfall and depth, and is mediated by genomically divergent microorganisms*, Nat Microbiol **4** (2019), no. 8, 1356–1367.

[98] S. Dickson, F. A. Smith, and S. E. Smith, *Structural differences in arbuscular mycorrhizal symbioses: more than 100 years after gallaud, where next?*, Mycorrhiza **17** (2007), no. 5, 375–393.

[99] W. Dietrich and O. Klimmek, *The function of methyl-menaquinone-6 and polysulfide reductase membrane anchor (psrc) in polysulfide respiration of wolinella succinogenes*, Eur J Biochem **269** (2002), no. 4, 1086–95.

[100] Michael S. Donnenberg, *Pathogenic strategies of enteric bacteria*, Nature **406** (2000), no. 6797, 768–774.

[101] W. F. Doolittle, *You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes*, Trends Genet. **14** (1998), 307–311.

[102] ———, *Lateral genomics*, Trends Cell. Biol. **9** (1999), M5–M8.

[103] Débora Farage Knupp dos Santos, Cynthia Maria Kyaw, Tatiana Amabile De Campos, Robert Neil Gerard Miller, Eliane Ferreira Noronha, Mercedes Maria da Cunha Bustamante, and Ricardo Kruger, *The family cystobacteraceae*, pp. 19–40, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

[104] D. F. R. Doud, R. M. Bowers, F. Schulz, M. De Raad, K. Deng, A. Tarver, E. Glasgow, K. V. Meulen, B. Fox, S. Deutsch, Y. Yoshikuni, T. Northen, B. P. Hedlund, S. W. Singer, N. Ivanova, and T. Woyke, *Function-driven single-cell genomics uncovers cellulose-degrading bacteria from the rare biosphere*, The ISME J. **14** (2019), 659–675.

[105] E. J .P. Douzery, E. A. Snell, E. Bapteste, F. Delsuc, and H. Philippe, *The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils?*, Proc. Nat. Acad. Sci. USA, **101** (2004), 15386–15391.

[106] I. S. Druzhinina, K. Chenthamara, J. Zhang, L. Atanasova, D. Yang, Y. Miao, M. J. Rahimi, M. Grujic, F. Cai, S. Pourmehdi, K. A. Salim, C. Pretzer, A. G. Kopchinskiy, B. Henrissat, A. Kuo, H. Hundley, M. Wang, A. Aerts, A. Salamov, A. Lipzen, K. LaButti, K. Barry, I. V. Grigoriev, Q. Shen, and C. P. Kubicek, *Massive lateral transfer of genes encoding plant cell wall-degrading enzymes to the mycoparasitic fungus trichoderma from its plant-associated hosts*, PLoS Genet **14** (2018), no. 4, e1007322.

[107] Xin-jun Du, Xiao-yi Wang, Xuan Dong, Ping Li, and Shuo Wang, *Characterization of the desiccation tolerance of cronobacter sakazakii strains*, Frontiers in Microbiology **9** (2018), no. 2867, NULL.

[108] I Duarte and MA. Huynen, *Contribution of lateral gene transfer to the evolution of the eukaryotic fungus piromyces sp. e2: Massive bacterial transfer of genes involved in carbohydrate metabolism.*, BioRxiv **514042** (2019), NULL.

[109] A. J. Dumbrell, M. Nelson, T. Helgason, C. Dytham, and A. H. Fitter, *Relative roles of niche and neutral processes in structuring a soil microbial community*, Isme j **4** (2010), no. 3, 337–45.

[110] M. Dworkin, *Nutritional requirements for vegetative growth of myxococcus xanthus*, J Bacteriol **84** (1962), no. 2, 250–7.

[111] L. Eichinger, J. A. Pachebat, G. Glockner, M. A. Rajandream, and et al., *The genome of the social amoeba dictyostelium discoideum*, Nature **435** (2005), 43–57.

[112] S. A. Eichorst, D. Trojan, S. Roux, C. Herbold, T. Rattei, and D. Woebken, *Genomic insights into the acidobacteria reveal strategies for their success in terrestrial environments*, Environ Microbiol **20** (2018), no. 3, 1041–1063.

[113] M. S. Elshahed, J. M. Senko, F. Z. Najar, S. M. Kenton, B. A. Roe, T. A. Dewers, J. R. Spear, and L. R. Krumholz, *Bacterial diversity and sulfur cycling in a mesophilic sulfide-rich spring*, Appl. Environ. Microbiol. **69** (2003), 5609–5621.

[114] B. Ely, *Genomic gc content drifts downward in most bacterial genomes*, PLoS One **16** (2021), no. 5, e0244163.

[115] L. Eme, E. Gentekaki, B. Curtis, J. M. Archibald, and A. J. Roger, *Lateral gene transfer in the adaptation of the anaerobic parasite blastocystis to the gut*, Curr Biol **27** (2017), no. 6, 807–820.

[116] H. A. Erikstad, S. Jensen, T. J. Keen, and N. K. Birkeland, *Differential expression of particulate methane monooxygenase genes in the verrucomicrobial methanotroph 'methylacidiphilum kamchatkense' kam1*, Extremophiles **16** (2012), no. 3, 405–9.

[117] K. F. Ettwig, M. K. Butler, D. Le Paslier, E. Pelletier, S. Mangenot, M. M. Kuypers, F. Schreiber, B. E. Dutilh, J. Zedelius, D. de Beer, J. Gloerich, H. J. Wessels, T. van Alen, F. Luesken, M. L. Wu, K. T. van de Pas-Schoonen, H. J. Op den Camp, E. M. Janssen-Megens, K. J. Francoijs, H. Stunnenberg, J. Weissenbach, M. S. Jetten, and M. Strous, *Nitrite-driven anaerobic methane oxidation by oxygenic bacteria*, Nature **464** (2010), no. 7288, 543–8.

[118] K. F. Ettwig, T. van Alen, K. T. van de Pas-Schoonen, M. S. Jetten, and M. Strous, *Enrichment and molecular detection of denitrifying methanotrophic bacteria of the nc10 phylum*, Appl Environ Microbiol **75** (2009), no. 11, 3656–62.

[119] I. F. Farag, J. P. Davis, N. H. Youssef, and M. S. Elshahed, *Global patterns of abundance, diversity and community structure of the aminicenantes (candidate phylum op8)*, PLoS ONE **9** (2014), e92139.

[120] I. Feussner and A. Polle, *What the transcriptome does not tell - proteomics and metabolomics are closer to the plants' patho-phenotype*, Curr Opin Plant Biol **26** (2015), 26–31.

[121] J. Fiedor, A. Sulikowska, A. Orzechowska, L. Fiedor, and K. Burda, *Antioxidant effects of carotenoids in a model pigment-protein complex*, Acta Biochim Pol **59** (2012), no. 1, 61–4.

[122] D. A. Fike, A. S. Bradley, and C. V. Rose, *Rethinking the anceint sulfur cycle*, Annu. Rev. Earth Planet. Sci. **43** (2015), 593–622.

[123] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. Sonnhammer, J. Tate, and M. Punta, *Pfam: the protein families database*, Nucleic Acids Res **42** (2014), no. Database issue, D222–30.

[124] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman, *The pfam protein families database: towards a more sustainable future*, Nucleic Acids Research **44** (2016), D279–D285.

[125] O. S. Fisher, G. E. Kenney, M. O. Ross, S. Y. Ro, B. E. Lemma, S. Batelu, P. M. Thomas, V. C. Sosnowski, C. J. DeHart, N. L. Kelleher, T. L. Stemmler, B. M. Hoffman, and A. C. Rosenzweig, *Characterization of a long overlooked copper protein from methane- and ammonia-oxidizing bacteria*, Nat Commun **9** (2018), no. 1, 4276.

[126] D. A. Fitzpatrick, *Horizontal gene transfer in fungi*, FEMS Microbiol. Lett. **2011** (2012), 1–8.

[127] V. Gadkar, J. D. Driver, and M. C. Rillig, *A novel in vitro cultivation system to produce and isolate soluble factors released from hyphae of arbuscular mycorrhizal fungi*, Biotechnol Lett **28** (2006), no. 14, 1071–6.

[128] V. Gadkar and M. C. Rillig, *The arbuscular mycorrhizal fungal protein glomalin is a putative homolog of heat shock protein 60*, FEMS Microbiol Lett **263** (2006), no. 1, 93–101.

[129] Xinliu Gao, Yueyong Xin, Patrick D. Bell, Jianzhong Wen, and Robert E. Blankenship, *Structural analysis of alternative complex iii in the photosynthetic electron transfer chain of chloroflexus aurantiacus*, Biochemistry **49** (2010), no. 31, 6670–6679.

[130] A. I. Garber, K. H Nealson, A. Okamoto, S. M. McAllister, C. S Chan, R. A. Barco, and N. Merino, *Fegenie: A comprehensive tool for the identification of iron genes and iron gene neighborhoods in genome and metagenome assemblies*, Front. Microbiol. **11** (2020), 37.

[131] Ronald Garcia and Rolf Müller, *The family haliangiaceae*, pp. 173–181, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

[132] _____, *The family myxococcaceae*, pp. 191–212, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

[133] _____, *The family nannocystaceae*, pp. 213–229, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

[134] Sarahi L. Garcia, Moritz Buck, Katherine D. McMahon, Hans-Peter Grossart, Alexander Eiler, and Falk Warnecke, *Auxotrophy and intrapopulation complementary in the 'interactome' of a cultivated freshwater model community*, Molecular Ecology **24** (2015), no. 17, 4449–4459.

[135] S. Garcia-Vallvé, A. Romeu, and J . Palau, *Horizontal gene transfer of glycosyl hydrolases of the rumen fungi*, Mol. Biol. Evol. **17** (2000), 352–361.

[136] Elisabeth Gasteiger, Christine Hoogland, Alexandre Gattiker, S'everine Duvaud, Marc R. Wilkins, Ron D. Appel, and Amos Bairoch, *Protein identification and analysis tools on the expasy server*, pp. 571–607, Humana Press, Totowa, NJ, 2005.

[137] E. A. Gladyshev, M. Meselson, and I. R. Arkhipova, *Massive horizontal gene transfer in bdelloid rotifers*, Science **320** (2008), 1210–1213.

[138] A. Gollotte, D. Van Tuinen, and D. Atkinson, *Diversity of arbuscular mycorrhizal fungi colonising roots of the grass species agrostis capillaris and lolium perenne in a field experiment*, Mycorrhiza **14** (2004), no. 2, 111–7.

[139] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev, *Full-length transcriptome assembly from rna-seq data without a reference genome.*, Nat. Biotechnol. **29** (2011), 644–652.

[140] David E. Graham, Marion Graupner, Huimin Xu, and Robert H. White, *Identification of coenzyme m biosynthetic 2-phosphosulfolactate phosphatase.* , European Journal of Biochemistry **268** (2001), no. 19, 5176–5188.

[141] J. R. Grant and L. A. Katz, *Phylogenomic study indicates widespread lateral gene transfer in entamoeba and suggests a past intimate relationship with parabasalids*, Genome Biol Evol **6** (2014), no. 9, 2350–60.

[142] A. G. Greene, *The family desulfuromonadaceae*, p. NULL, Springer-Verlag, Berlin Heidelberg, 2014.

[143] R. J. Gruninger, A. K. Puniyab, T. M. Callaghanc, J. E. Edwardsc, Noha Youssef, S. S. Dagare, K. Fliegerova, G. W. Griffith, R. Forster, A. Tsang, Tim McAllister, and M. S. Elshahed, *Anaerobic fungi (phylum neocallimastigomycota): Advances in understanding of their taxonomy, life cycle, ecology, role, and biotechnological potential*, FEMS Microbiol. Ecol. **Under Review** (2014), NULL.

[144] N. Gupta and P. A. Pevzner, *False discovery rates of protein identifications: a strike against the two-peptide rule*, J Proteome Res **8** (2009), no. 9, 4173–81.

[145] B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, MB Couger, D. Eccles, B. Li, M. Lieber, M. D MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N Dewey, R. Henschel, R. D LeDuc, N. Friedman, and A. Regev, *De novo transcript sequence reconstruction from rna-seq using the trinity platform for reference generation and analysis*, Nat. Protocols **8** (2013), 1494–1512.

[146] J. H. Hackstein and N. Yarlett, *Hydrogenosomes and symbiosis*, Prog Mol Subcell Biol **41** (2006), 117–42.

[147] M. W. Hahn, J. Schmidt, U. Koll, M. Rohde, S. Verbarg, A. Pitt, R. Nakai, T. Naganuma, and E. Lang, *Silvanigrella aquatica gen. nov., sp. nov., isolated from a freshwater lake, description of silvanigrellaceae fam. nov. and silvanigrellales ord. nov., reclassification of the order bdellovibrionales in the class oligoflexia, reclassification of the families bacteriovoracaceae and halobacteriovoraceae in the new order bacteriovoracales ord. nov., and reclassification of the family pseudobacteriovoracaceae in the order oligoflexales*, Int. J. Syst. Evoll. Microbiol. **67** (2017), 2555–2568.

[148] C. H. Haitjema, S. P. Gilmore, J. K. Henske, K. V. Solomon, R. de Groot, A. Kuo, S. J. Mondo, A. A. Salamov, K. LaButti, Z. Zhao, J. Chiniquy, K. Barry, H. M. Brewer, S. O. Purvine, A. T. Wright, M. Hainaut, B. Boxma, T. van Alen, J. H. P. Hackstein, B. Henrissat, S. E. Baker, I. V. Grigoriev, and M. A. O'Malley, *A parts list for fungal cellulosomes revealed by comparative genomics*, Nature Microbiol. **2** (2017), 17087.

[149] N. Hamamura, C. M. Yeager, and D. J. Arp, *Two distinct monooxygenases for alkane oxidation in nocardioides sp. strain cf8*, Appl Environ Microbiol **67** (2001), no. 11, 4992–8.

[150] R. A. Hanafy, M. S. Elshahed, A. S. Liggenstoffer, G. W. Griffith, and N. H. Youssef, *Pecoramyces ruminantium, gen. nov., sp. nov., an anaerobic gut fungus from the feces of cattle and sheep*, Mycologia **109** (2017), 231–243.

[151] R. A. Hanafy, M. S. Elshahed, and N. H. Youssef, *Feramyces austinii, gen. nov, sp. nov., an anaerobic gut fungus from rumen and fecal samples of wild barbary sheep and fallow deer*, Submitted (2018), NULL.

[152] H. R. Harhangi, A. S. Akhmanova, R. Emmens, C. van der Drift, W. T. A. M. de Laat, J. P. van Dijken, M. S. M. Jetten, J. T. Pronk, and H. J. M. Op den Camp, *Xylose metabolism in the anaerobic fungus piromyces sp. strain e2 follows the bacterial pathway*, Arch. Microbiol. **180** (2003), 134–142.

[153] H. Hashimoto, C. Uragami, and R. J. Cogdell, *Carotenoids and photosynthesis*, Subcell Biochem **79** (2016), 111–39.

[154] Saad El-Din Hassan, Terrence H. Bell, Franck O. P. Stefani, David Denis, Mohamed Hijri, and Marc St-Arnaud, *Contrasting the community structure of arbuscular mycorrhizal fungi from hydrocarbon-contaminated and uncontaminated soils following willow (salix spp. l.) planting*, PLOS ONE **9** (2014), no. 7, e102838.

[155] Christine V. Hawkes, Iain P. Hartley, Phil Ineson, and Alastair H. Fitter, *Soil temperature affects carbon allocation within arbuscular mycorrhizal networks and carbon transport from plant to fungus*, Global Change Biology **14** (2008), no. 5, 1181–1190.

[156] P. A. Haynes and T. H. Roberts, *Subcellular shotgun proteomics in plants: looking beyond the usual suspects*, Proteomics **7** (2007), no. 16, 2963–75.

[157] J. He, L. Yi, L. Hai, L. Ming, W. Gao, and R. Ji, *Characterizing the bacterial microbiota in different gastrointestinal tract segments of the bactrian camel*, Sci Rep **8** (2018), no. 1, 654.

[158] Shaomei He, Stephanie A. Malfatti, Jack W. McFarland, Frank E. Anderson, Amrita Pati, Marcel Huntemann, Julien Tremblay, Tijana Glavina del Rio, Mark P. Waldrop, Lisamarie Windham-Myers, and Susannah G. Tringe, *Patterns in wetland microbial community composition and functional gene repertoire associated with methane emissions*, mBio **6** (2015), no. 3, e00066–15.

[159] A. Heinemeyer and A. H. Fitter, *Impact of temperature on the arbuscular mycorrhizal (am) symbiosis: growth responses of the host plant and its am fungal partner*, J Exp Bot **55** (2004), no. 396, 525–34.

[160] Harm J. Hektor, Harm Kloosterman, and Lubbert Dijkhuizen, *Nicotinoprotein methanol dehydrogenase enzymes in gram-positive methylotrophic bacteria*, Journal of Molecular Catalysis B: Enzymatic **8** (2000), no. 1, 103–109.

[161] T. Helgason, A. H. Fitter, and J. P. W. Young, *Molecular diversity of arbuscular mycorrhizal fungi colonising hyacinthoides non-scripta (bluebell) in a seminatural woodland*, Molecular Ecology **8** (1999), no. 4, 659–666.

[162] F. Hildebrand, A. Meyer, and A. Eyre-Walker, *Evidence of selection upon genomic gc-content in bacteria*, PLoS Genet **6** (2010), no. 9, e1001107.

[163] M. Hinojosa-Leon, M. Dubourdieu, J. A. Sanchez-Crispin, and M. Chippaux, *Tetrathionate reductase of salmonella thyphimurium: a molybdenum containing enzyme*, Biochem Biophys Res Commun **136** (1986), no. 2, 577–81.

[164] R. P. Hirt, N. Harriman, A. V. Kajava, and T. M. Embley, *A novel potential surface protein in trichomonas vaginalis contains a leucine-rich repeat shared by microorganisms from all three domains of life.*, Mol. Biochem. Parasitol. **125** (2002), 195–199.

[165] Y. W. Ho and D J S Barr, *Classification of anaerobic gut fungi from herbivores with emphasis on rumen fungi from malaysia*, mycologia **87** (1995), no. 5, 655–677.

[166] Pierre Hohmann and Monika M. Messmer, *Breeding for mycorrhizal symbiosis: focus on disease resistance*, Euphytica **213** (2017), no. 5, 113.

[167] M. F. Hohmann-Marriott and R. E. Blankenship, *Evolution of photosynthesis*, Annu Rev Plant Biol **62** (2011), 515–48.

[168] S. Horn, S. Hempel, E. Verbruggen, M. C. Rillig, and T. Caruso, *Linking the community structure of arbuscular mycorrhizal fungi and plants: a story of interdependence?*, Isme j **11** (2017), no. 6, 1400–1411.

[169] P. Hu, A. E. Dubinsky, A. J. Probst, J. Wang, C. M. K. Sieber, L. M. Tom, P. R. Gardinali, J. F. Banfield, R. M. Atllas, and G. L. Andersen, *Simulation of deepwater horizon oil plume reveals substrate specialization within a complex community of hydrocarbon degraders*, Proc. Natl. Acad. Sci. USA **114** (2017), 7432–7437.

[170] J. L. Huang, *Horizontal gene transfer in eukaryotes: the weak-link model*, Bioassays **35** (2013), 868–875.

[171] L. Huang, H. Zhang, P. Wu, S. Entwistle, X. Li, T. Yohe, H. Yi, Z. Yang, and Y. Yin, *dbcan-seq: a database of carbohydrate-active enzyme (cazyme) sequence and annotation*, Nucleic Acids Res **46** (2018), no. D1, D516–D521.

[172] N. C. Hubner, S. Ren, and M. Mann, *Peptide separation with immobilized pi strips is an attractive alternative to in-gel protein digestion for proteome analysis*, Proteomics **8** (2008), no. 23-24, 4862–72.

[173] Laura A Hug, Brett J Baker, Karthik Anantharaman, Christopher T Brown, Alexander J Probst, Cindy J Castelle, Cristina N Butterfield, Alex W Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A Relman, Kari M Finstad, Ronald Amundson, Brian C Thomas, and Jillian F Banfield, *A new view of the tree of life*, Nature Microbiology (2016/4/11), 16048.

[174] Laura A. Hug and Rebecca Co, *It takes a village: Microbial communities thrive through interactions and metabolic handoffs*, mSystems **3** (2018), no. 2, e00152–17.

[175] R. E. Hungate, *A roll tube method for cultivation of strict anaerobes*, Meth. Microbiol. **3** (1969), 117–132.

[176] S. Huntley, K. Wuichet, and L. Søgaard-Andersen, *Genome evolution and content in the myxobacteria*, vol. 1, pp. 31–50, Caister Academic Press, 2014.

[177] R. Husband, E. A. Herre, S. L. Turner, R. Gallery, and J. P. Young, *Molecular diversity of arbuscular mycorrhizal fungi and patterns of host association over time and space in a tropical forest*, Mol Ecol **11** (2002), no. 12, 2669–78.

[178] F. Husnik and J. P. McCutcheon, *Functional horizontal gene transfer from bacteria to eukaryotes*, Nat. Rev. Microbiol. **doi:10.1038/nrmicro.2017.137** (2017), NULL.

[179] D. Hyatt, G-L Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J Hauser, *Prodigal: prokaryotic gene recognition and translation initiation site identification*, BMC Bioinformatics **11** (2010), 119.

[180] H. Iguchi, H. Yurimoto, and Y. Sakai, *Soluble and particulate methane monooxygenase gene clusters of the type i methanotroph methylovulum miyakonense ht12*, FEMS Microbiol Lett **312** (2010), no. 1, 71–6.

[181] Takashi Iizuka, Mitsunori Tokura, Yasuko Jojima, Akira Hiraishi, Shigeru Yamanaka, and Ryosuke Fudou, *Enrichment and phylogenetic analysis of moderately thermophilic myxobacteria from hot springs in japan*, Microbes and Environments **21** (2006), no. 3, 189–199.

[182] H. Innan and F. Kondrashov, *The evolution of gene duplications: classifying and distinguishing between models*, Nat. Rev. Genet. **11** (2010), 97–10.

[183] I. C. Irvine, C. A. Brigham, K. N. Suding, and J. B. Martiny, *The abundance of pink-pigmented facultative methylotrophs in the root zone of plant species in invaded coastal sage scrub habitat*, PLoS One **7** (2012), no. 2, e31026.

[184] B. E. Jackson, V. K. Bhupathiraju, R. S. Tanner, C. R. Woese, and M. J. McInerney, *Syntrophus aciditrophicus sp. nov., a new anaerobic bacterium that degrades fatty acids and benzoate in syntrophic association with hydrogen-using microorganisms*, Arch Microbiol **171** (1999), 107–114.

[185] Iver Jakobsen, *Hyphal fusion to plant species connections – giant mycelia and community nutrient flow*, New Phytologist **164** (2004), no. 1, 4–7.

[186] Jan Jansa, Ahmad Mozafar, and Emmanuel Frossard, *Long-distance transport of p and zn through the hyphae of an arbuscular mycorrhizal fungus in symbiosis with maize*, Agronomie **23** (2003), no. 5-6, 481–488.

[187] Vinicio D. Armijos Jaramillo, Serenella A. Sukno, and Michael R. Thon, *Identification of horizontally transferred genes in the genus colletotrichum reveals a steady tempo of bacterial to fungal gene transfer*, BMC Genomics **16** (2015), no. 1, 2.

[188] Y. Ji, R. Angel, M. Klose, P. Claus, H. Marotta, L. Pinho, A. Enrich-Prast, and R. Conrad, *Structure and function of methanogenic microbial communities in sediments of amazonian lakes with different water types*, Environ Microbiol **18** (2016), no. 12, 5082–5100.

[189] Melanie D. Jones and Sally E. Smith, *Exploring functional definitions of mycorrhizas: Are mycorrhizas always mutualisms?*, Canadian Journal of Botany **82** (2004), no. 8, 1089–1109.

[190] A. Jordan and P. Reichard, *Ribonucleotide reductases*, Annual Review of Biochemistry **67** (1998), no. 1, 71–98.

[191] C. Jégousse, P. Vannier, R. Groben, F. O. Glöckner, and V. Marteinsson, *A total of 219 metagenome-assembled genomes of microorganisms from icelandic marine waters*, PeerJ **9** (2021), e11112.

[192] H. Kaessmann, *Origins, evolution, and phenotypic impact of new genes2*, Genome Res. **20** (2010), 1313–1326.

[193] M. G. Kalyuzhnaya, K. R. Hristova, M. E. Lidstrom, and L. Chistoserdova, *Characterization of a novel methanol dehydrogenase in representatives of burkholderiales: implications for environmental detection of methylotrophy and evidence for convergent evolution*, J Bacteriol **190** (2008), no. 11, 3817–23.

[194] M. G. Kalyuzhnaya, R. Zabinsky, S. Bowerman, D. R. Baker, M. E. Lidstrom, and L. Chistoserdova, *Fluorescence in situ hybridization-flow cytometry-cell sorting-based method for separation and enrichment of type i and type ii methanotroph populations*, Appl Environ Microbiol **72** (2006), no. 6, 4293–301.

[195] P. P. Kandel, Z. Pasternak, J. van Rijn, O. Nahum, and E. Jurkevitch, *Abundance, diversity and seasonal dynamics of predatory bacteria in aquaculture zero discharge systems*, FEMS Microbiol Ecol **89** (2014), no. 1, 149–61.

[196] M. Kanehisa and Y. Sato, *Kegg mapper for inferring cellular functions from protein sequences*, Protein Sci **29** (2020), no. 1, 28–35.

[197] M. Kanehisa, Y. Sato, and K. Morishima, *Blastkoala and ghostkoala: Kegg tools for functional characterization of genome and metagenome sequences*, J Mol Biol **428** (2016), no. 4, 726–731.

[198] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe, *Kegg for integration and interpretation of large-scale molecular data sets*, Nucleic acids research **40** (2012), no. D1, D109–D114.

[199] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe, *Kegg as a reference resource for gene and protein annotation*, Nucleic Acids Research **44** (2015), no. D1, D457–D462.

[200] D. D. Kang, F. Li, E. Kirton, A. Thomas, R. Egan, H. An, and Z. Wang, *Metabat 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies*, PeerJ **7** (2019), e7359.

[201] K. Katoh and D. M. Standley, *Mafft multiple sequence alignment software version 7: improvements in performance and usability*, Mol Biol Evol **30** (2013), no. 4, 772–80.

[202] P. J. Keeling and J.D. Palmer, *Horizontal gene transfer in eukaryotic evolution*, Nat. Rev. Genet. **9** (2008), 605–618.

[203] Valentina N. Khmelenina, J. Colin Murrell, Thomas J. Smith, and Yuri A. Trotsenko, *Physiology and biochemistry of the aerobic methanotrophs*, pp. 1–25, Springer International Publishing, Cham, 2018.

[204] S. P. Kishore, J. W. Stiller, and K. W. Deitsch, *Horizontal gene transfer of epigenetic machinery and evolution of parasitism in the malaria parasite plasmodium falciparum and other apicomplexans*, BMC Evol. Biol. **13** (2013), 37.

[205] C. Knief, *Diversity and habitat preferences of cultivated and uncultivated aerobic methanotrophic bacteria evaluated based on pmoa as molecular marker*, Front Microbiol **6** (2015), 1346.

[206] Y. Kobayashi, T. Maeda, K. Yamaguchi, H. Kameoka, S. Tanaka, T. Ezawa, S. Shigenobu, and M. Kawaguchi, *The genome of rhizophagus clarus hr1 reveals a common genetic basis for auxotrophy among arbuscular mycorrhizal fungi*, BMC Genomics **19** (2018), no. 1, 465.

[207] Steffen Kolb, *Aerobic methanol-oxidizing bacteria in soil*, FEMS Microbiology Letters **300** (2009), no. 1, 1–10.

[208] Konstantinos T. Konstantinidis, Ramon Rosselló-Móra, and Rudolf Amann, *Uncultivated microbes in need of their own taxonomy*, The ISME Journal **11** (2017), no. 11, 2399–2406.

[209] S. K. Kothari, H. Marschner, and V. RÖMheld, *Effect of a vesicular–arbuscular mycorrhizal fungus and rhizosphere micro-organisms on manganese reduction in the rhizosphere and manganese concentrations in maize (zea mays l.)*, New Phytologist **117** (1991), no. 4, 649–655.

[210] W. Kou, J. Zhang, X. Lu, Y. Ma, X. Mou, and L. Wu, *Identification of bacterial communities in sediments of poyang lake, the largest freshwater lake in china*, Springerplus **5** (2016), 401.

[211] G. Koutsovoulos, S. Kumar, D. R. Laetsch, L. Stevens, J. Daub, C. Conlon, H. Maroon, F. Thomas, A. A. Aboobaker, and M. Blaxter, *No evidence for extensive horizontal gene transfer in the genome of the tardigrade hypsibius dujardini*, Proc. Nat. Acad. Sci. USA **113** (2016), 5053–5058.

[212] A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, and A. Stamatakis, *Raxml-ng: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference*, Bioinformatics **35** (2019), no. 21, 4453–4455.

[213] Ramasamy Krishnamoorthy, Chang-Gi Kim, Parthiban Subramanian, Ki-Yoon Kim, Gopal Selvakumar, and Tong-Min Sa, *Arbuscular mycorrhizal fungi community structure, abundance and species richness changes in soil by different levels of heavy metal and metalloid concentration*, PLOS ONE **10** (2015), no. 6, e0128784.

[214] Anita Krisko and Miroslav Radman, *Biology of extreme radiation resistance: The way of deinococcus radiodurans*, Cold Spring Harbor Perspectives in Biology **5** (2013), no. 7, NULL.

[215] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer, *Predicting transmembrane protein topology with a hidden markov model: application to complete genomes*, J Mol Biol **305** (2001), no. 3, 567–80.

[216] Lee Kroos, *Highly signal-responsive gene regulatory network governing myxococcus development*, Trends in Genetics **33** (2017), no. 1, 3–15.

[217] M. Kruger, C. Kruger, C. Walker, H. Stockinger, and A. Schussler, *Phylogenetic reference data for systematics and phylotaxonomy of arbuscular mycorrhizal fungi from phylum to species level*, New Phytol **193** (2012), no. 4, 970–84.

[218] J. G. Krum and S. A. Ensign, *Heterologous expression of bacterial epoxyalkane:coenzyme m transferase and inducible coenzyme m biosynthesis in xanthobacter strain py2 and rhodococcus rhodochrous b276*, J Bacteriol **182** (2000), no. 9, 2629–34.

[219] J. Kuever, *The family desulfobulbaceae*, p. NULL, Springer-Verlag, Berlin Heidelberg, 2014.

[220] K. M. Kuivila, J. W. Murray, A. H. Devol, M. E. Lidstrom, and C. E. Reimers, *Methane cycling in the sediments of lake washington*, Limnology and Oceanography **33** (1988), no. 4, 571–581.

[221] Sudhir Kumar, Glen Stecher, Michael Li, Christina Knyaz, and Koichiro Tamura, *Mega x: Molecular evolutionary genetics analysis across computing platforms*, Molecular Biology and Evolution **35** (2018), no. 6, 1547–1549.

[222] K. Lagesen, P. Hallin, E. A. Rødland, H. H. Staerfeldt, T. Rognes, and D. W. Ussery, *Rnammer: consistent and rapid annotation of ribosomal rna genes*, Nucleic Acids Res **35** (2007), no. 9, 3100–8.

[223] K. Lagesen, P. Hallin, E. A. Rødland, H-H Staerfeldt, T .Rognes, and D. W Ussery, *Rnammer: consistent and rapid annotation of ribosomal rna genes*, Nucleic Acids Res **35** (2007), 3100–3108.

[224] B. Langmead and S. L. Salzberg, *Fast gapped-read alignment with bowtie 2*, Nat. Methods **9** (2012), 357–359.

[225] F. Lassalle, S. Périan, T. Bataillon, X. Nesme, L. Duret, and V. Daubin, *Gc-content evolution in bacterial genomes: the biased gene conversion hypothesis expands*, PLoS Genet **11** (2015), no. 2, e1004941.

[226] Jean Le Mer and Pierre Roger, *Production, oxidation, emission and consumption of methane by soils: A review*, European Journal of Soil Biology **37** (2001), no. 1, 25–50.

[227] Changhan Lee and Chankyu Park, *Bacterial responses to glyoxal and methylglyoxal: Reactive electrophilic species*, International Journal of Molecular Sciences **18** (2017), no. 1, 169.

[228] J. Lee, S. Lee, and J. P. Young, *Improved pcr primers for the detection and identification of arbuscular mycorrhizal fungi*, FEMS Microbiol Ecol **65** (2008), no. 2, 339–49.

[229] Michelle M. Leger, Laura Eme, Laura A. Hug, and Andrew J. Roger, *Novel hydrogenosomes in the microaerophilic jakobid stygiella incarcerata*, Molecular Biology and Evolution **33** (2016), no. 9, 2318–2336.

[230] Ylva Lekberg, Roger T. Koide, Jason R. Rohr, Laura Aldrich-Wolfe, and Joseph B. Morton, *Role of niche restrictions and dispersal in the composition of arbuscular mycorrhizal fungal communities*, Journal of Ecology **95** (2007), no. 1, 95–105.

[231] Ivica Letunic and Peer Bork, *Interactive tree of life (itol) v3: an online tool for the display and annotation of phylogenetic and other trees*, Nucleic acids research **44** (2016), no. W1, W242–W245.

[232] D. Li, C-M. Liu, R. Luo, K. Sadakane, and T-W Lam, *Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph*, Bioinformatics **31** (2015), 1674–1676.

[233] L. Li, X. Liu, W. Yang, F. Xu, W. Wang, L. Feng, M. Bartlam, L. Wang, and Z. Rao, *Crystal structure of long-chain alkane monooxygenase (lada) in complex with coenzyme fmn: unveiling the long-chain alkane hydroxylase*, J Mol Biol **376** (2008), no. 2, 453–65.

[234] M. Li, S. Jain, B. J. Baker, C. Taylor, and G. J. Dick, *Novel hydrocarbon monooxygenase genes in the metatranscriptome of a natural deep-sea hydrocarbon plume*, Environ Microbiol **16** (2014), no. 1, 60–71.

[235] S. G. Li, X. W. Zhou, P. F. Li, K. Han, W. Li, Z. F. Li, Z. H. Wu, and Y. Z. Li, *The existence and diversity of myxobacteria in lake mud - a previously unexplored myxobacteria habitat*, Environ Microbiol Rep **4** (2012), no. 6, 587–95.

[236] Huawen Lin, Alan L. Kwan, and Susan K. Dutcher, *Synthesizing and salvaging nad+: Lessons learned from chlamydomonas reinhardtii*, PLOS Genetics **6** (2010), no. 9, e1001105.

[237] Kui Lin, Erik Limpens, Zhonghua Zhang, Sergey Ivanov, Diane G. O. Saunders, Desheng Mu, Erli Pang, Huifen Cao, Hwangho Cha, Tao Lin, Qian Zhou, Yi Shang, Ying Li, Trupti Sharma, Robin van Velzen, Norbert de Ruijter, Duur K. Aanen, Joe Win, Sophien Kamoun, Ton Bisseling, René Geurts, and Sanwen Huang, *Single nucleus genome sequencing reveals high similarity among nuclei of an endomycorrhizal fungus*, PLOS Genetics **10** (2014), no. 1, e1004078.

[238] X. Lin, Y. Feng, H. Zhang, R. Chen, J. Wang, J. Zhang, and H. Chu, *Long-term balanced fertilization decreases arbuscular mycorrhizal fungal diversity in an arable soil in north china revealed by 454 pyrosequencing*, Environ Sci Technol **46** (2012), no. 11, 5764–71.

[239] Rui Liu, Ruining Cai, Jing Zhang, and Chaomin Sun, *Heimdallarchaeota harness light energy through photosynthesis*, bioRxiv (2020), 2020.02.20.957134.

[240] Y. Liu, D. L. Balkwill, H. C. Aldrich, G. R. Drake, and D. R. Boone, *Characterization of the anaerobic propionate-degrading syntrophs smithella propionica gen. nov, sp. nov. and syntrophobacter wolinii*, Int. J. Syst. Bacteriol. **49** (1999), 545–556.

[241] Marc Llirós, Jean-Pierre Descy, Xavier Libert, Cédric Morana, Mélodie Schmitz, Louisette Wimba, Angélique Nzavuga-Izere, Tamara García-Armisen, Carles Borrego, Pierre Servais, and François Darchambeau, *Microbial ecology of lake kivu*, pp. 85–105, Springer Netherlands, Dordrecht, 2012.

[242] D. R. Lovley, S. J. Giovannoni, D. C. White, J. E. Champine, E. J. Phillips, Y. A. Gorby, and S. Goodwin, *Geobacter metallireducens gen. nov. sp. nov., a microorganism capable of coupling the complete oxidation of organic compounds to the reduction of iron and other metals*, Arch Microbiol **159** (1993), 336–344.

[243] T. Maeda, Y. Kobayashi, H. Kameoka, N. Okuma, N. Takeda, K. Yamaguchi, T. Bino, S. Shigenobu, and M. Kawaguchi, *Evidence of non-tandemly repeated rdnas and their intragenomic heterogeneity in rhizophagus irregularis*, Commun Biol **1** (2018), 87.

[244] M. Magot, G. Ravot, X. Campaignolle, B. Ollivier, B. K. Patel, M. L. Fardeau, P. Thomas, J. L. Crolet, and J. L. Garcia, *Dethiosulfovibrio peptidovorans gen. nov., sp. nov., a new anaerobic, slightly halophilic, thiosulfate-reducing bacterium from corroding offshore oil wells*, Int J Syst Bacteriol **47** (1997), no. 3, 818–24.

[245] M. Marcet-Bouben and T. Gabaldon, *Acquisition of prokaryotic genes by fungal genomes*, Trends Genet. **26** (2010), 5–8.

[246] A. Marchler-Bauer, Y. Bo, L. Han, J. He, C. J. Lanczycki, S. Lu, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, F. Lu, G. H. Marchler, J. S. Song, N. Thanki, Z. Wang, R. A. Yamashita, D. Zhang, C. Zheng, L. Y. Geer, and S. H. Bryant, *Cdd/sparcle: functional classification of proteins via subfamily domain architectures*, Nucleic Acids Res **45** (2017), no. D1, D200–d203.

[247] R. C. Marshall and D. E. Whitworth, *Is "wolf-pack" predation by antimicrobial bacteria cooperative? cell behaviour and predatory mechanisms indicate profound selfishness, even when working alongside kin*, Bioessays **41** (2019), no. 4, e1800247.

[248] W. F. Martin, D. A. Bryant, and J. T. Beatty, *A physiological perspective on the origin and evolution of photosynthesis*, FEMS Microbiol Rev **42** (2018), no. 2, 205–231.

[249] C. G. P. McCarthy and D. A. Fitzpatrick, *Systematic search for evidence of interdomain horizontal gene transfer from prokaryotes to oomycete lineages*, mSphere **1** (2016), e00195–16.

[250] Thomas M. McCollom, *Laboratory simulations of abiotic hydrocarbon formation in earth's deep subsurface*, Reviews in Mineralogy and Geochemistry **75** (2013), no. 1, 467–494.

[251] T. L. McTaggart, D. A. Beck, U. Setboonsarng, N. Shapiro, T. Woyke, M. E. Lidstrom, M. G. Kalyuzhnaya, and L. Chistoserdova, *Genomics of methylotrophy in gram-positive methylamine-utilizing bacteria*, Microorganisms **3** (2015), no. 1, 94–112.

[252] M. H. Medema, K. Blin, P. Cimermancic, V. de Jager, P. Zakrzewski, M. A. Fischbach, T. Weber, E. Takano, and R. Breitling, *antismash: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences*, Nucleic Acids Res **39** (2011), no. Web Server issue, W339–46.

[253] K. Mendler, H. Chen, D. H. Parks, B. Lobb, L. A. Hug, and A. C. Doxey, *Annotree: visualization and exploration of a functionally annotated microbial tree of life*, Nucleic Acids Res **47** (2019), no. 9, 4442–4448.

[254] Jun Meng, Fengping Wang, Feng Wang, Yanping Zheng, Xiaotong Peng, Huaiyang Zhou, and Xiang Xiao, *An uncultivated crenarchaeota contains functional bacteriochlorophyll a synthase*, The ISME Journal **3** (2009), no. 1, 106–116.

[255] R. M. Miller and J. D. Jastrow, *Hierarchy of root and mycorrhizal fungal interactions with soil aggregation*, Soil Biology and Biochemistry **22** (1990), no. 5, 579–584.

[256] _____, *Mycorrhizal fungi influence soil structure*, pp. 3–18, Springer Netherlands, Dordrecht, 2000.

[257] B. Q. Minh, M. A. Nguyen, and A. von Haeseler, *Ultrafast approximation for phylogenetic bootstrap*, Mol Biol Evol **30** (2013), no. 5, 1188–95.

[258] Jaina Mistry, Robert D. Finn, Sean R. Eddy, Alex Bateman, and Marco Punta, *Challenges in homology search: Hmmer3 and convergent evolution of coiled-coil regions*, Nucleic Acids Research **41** (2013), no. 12, e121–e121.

[259] S. Mittal and L. Kroos, *Combinatorial regulation by a novel arrangement of frua and mrpc2 transcription factors during myxococcus xanthus development*, J Bacteriol **191** (2009), no. 8, 2753–63.

[260] Daniel J. Moebius-Clune, Bianca N. Moebius-Clune, Harold M. van Es, and Teresa E. Pawlowska, *Arbuscular mycorrhizal fungi associated with a single agronomic plant host across the landscape: Community differentiation along a soil textural gradient*, Soil Biology and Biochemistry **64** (2013), 191–199.

[261] Sepehr S. Mohammadi, Arjan Pol, Theo van Alen, Mike S. M. Jetten, and Huub J. M. Op den Camp, *Ammonia oxidation and nitrite reduction in the verrucomicrobial methanotroph methylacidiphilum fumariolicum solv*, Frontiers in Microbiology **8** (2017), no. 1901, NULL.

[262] Kathrin I. Mohr, *Diversity of myxobacteria-we only see the tip of the iceberg*, Microorganisms **6** (2018), no. 3, 84.

[263] C. Moliner, P. E. Fournier, and D. Raoult, *Genome analysis of microorganisms living in amoebae reveals a melting pot of evolution*, FEMS Microbiol Rev **34** (2010), no. 3, 281–94.

[264] Emmanuelle Morin, Shingo Miyauchi, Hélène San Clemente, Eric C. H. Chen, Adrian Pelin, Ivan de la Providencia, Steve Ndikumana, Denis Beaudet, Mathieu Hainaut, Elodie Drula, Alan Kuo, Nianwu Tang, Sébastien Roy, Julie Viala, Bernard Henrissat, Igor V. Grigoriev, Nicolas Corradi, Christophe Roux, and Francis M. Martin, *Comparative genomics of rhizophagus irregularis, r. cerebriforme, r. diaphanus and gigaspora rosea highlights specific genetic features in glomeromycotina*, New Phytologist **222** (2019), no. 3, 1584–1598.

[265] C. G. Mowat, E. Rothery, C. S. Miles, L. McIver, M. K. Doherty, K. Drewette, P. Taylor, M. D. Walkinshaw, S. K. Chapman, and G. A. Reid, *Octaheme tetrathionate reductase is a respiratory enzyme with novel heme ligation*, Nat Struct Mol Biol **11** (2004), no. 10, 1023–4.

[266] S. Mukherjee, D. Stamatis, J. Bertsch, G. Ovchinnikova, H. Y. Katta, A. Mojica, I-M. A. Chen, N. C. Kyrpides, and T. B. K. Reddy, *Genomes online database (gold) v.7: updates and new features*, Nucl. Acids Res. **47** (2019), D649–D659.

[267] Chelsea L. Murphy, James Biggerstaff, Alexis Eichhorn, Essences Ewing, Ryan Shahan, Diana Soriano, Sydney Stewart, Kaitlynn VanMol, Ross Walker, Payton Walters, Mostafa S. Elshahed, and Noha H. Youssef, *Genomic characterization of three novel desulfobacterota classes expand the metabolic and phylogenetic diversity of the phylum*, Environmental Microbiology **n/a** (2021), no. n/a, NULL.

[268] Chelsea L. Murphy, Peter F. Dunfield, Andriy Sheremet, John R. Spear, Ramunas Stepanauskas, Tanja Woyke, Mostafa S. Elshahed, and Noha H. Youssef, *Methylotrophy, alkane-degradation, and pigment production as defining features of the globally distributed yet-uncultured phylum binatota*, bioRxiv (2020), 2020.09.14.296780.

[269] H. Musto, H. Naya, A. Zavala, H. Romero, F. Alvarez-Valín, and G. Bernardi, *Genomic gc level, optimal growth temperature, and genome size in prokaryotes*, Biochem Biophys Res Commun **347** (2006), no. 1, 1–3.

[270] G. Muyzer and A. J. Stams, *The ecology and biotechnology of sulphate-reducing bacteria*, Nat Rev Microbiol **6** (2008), no. 6, 441–54.

[271] J. Muñoz-Dorado, A. Moraleda-Muñoz, F. J. Marcos-Torres, F. J. Contreras-Moreno, A. B. Martin-Cuadrado, J. M. Schrader, P. I. Higgs, and J. Pérez, *Transcriptome dynamics of the myxococcus xanthus multicellular developmental program*, Elife **8** (2019), NULL.

[272] José Muñoz-Dorado, Francisco J. Marcos-Torres, Elena García-Bravo, Aurelio Moraleda-Muñoz, and Juana Pérez, *Myxobacteria: Moving, killing, feeding, and surviving together*, Frontiers in Microbiology **7** (2016), no. 781, NULL.

[273] F. D. Müller, C. W. Schink, E. Hoiczyk, E. Cserti, and P. I. Higgs, *Spore formation in myxococcus xanthus is tied to cytoskeleton functions and polysaccharide spore coat deposition*, Mol Microbiol **83** (2012), no. 3, 486–505.

[274] F. D. Müller, A. Treuner-Lange, J. Heider, S. M. Huntley, and P. I. Higgs, *Global transcriptome analysis of spore formation in myxococcus xanthus reveals a locus necessary for cell differentiation*, BMC Genomics **11** (2010), 264.

[275] V. Müller, *Energy conservation in acetogenic bacteria*, Appl Environ Microbiol **69** (2003), no. 11, 6345–53.

[276] Y. Nagata, K. Miyauchi, J. Damborsky, K. Manova, A. Ansorgova, and M. Takagi, *Purification and characterization of a haloalkane dehalogenase of a new substrate class from a gamma-hexachlorocyclohexane-degrading bacterium, sphingomonas paucimobilis ut26*, Appl Environ Microbiol **63** (1997), no. 9, 3707–10.

[277] K. Nakai and P. Horton, *Psort: a program for detecting sorting signals in proteins and predicting their subcellular localization*, Trends Biochem Sci **24** (1999), no. 1, 34–6.

[278] T. Nakamura, K. D. Yamada, K. Tomii, and K. Katoh, *Parallelization of mafft for large-scale multiple sequence alignments*, Bioinformatics **34** (2018), no. 14, 2490–2492.

[279] H. Naya, H. Romero, A. Zavala, B. Alvarez, and H. Musto, *Aerobiosis increases the genomic guanine plus cytosine content (gc%) in prokaryotes*, J Mol Evol **55** (2002), no. 3, 260–4.

[280] S. Nayfach, S. Roux, R. Seshadri, D. Udwary, N. Varghese, F. Schulz, and et al., *A genomic catalogue of earth's microbiomes*, Nat Biotechnol **Accpeted** (2020), NULL.

[281] Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh, *Iq-tree: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies*, Molecular Biology and Evolution **32** (2015), no. 1, 268–274.

[282] J. E. J. Nixon, A. Wang, J. Field, H. G. Morrison, A. G. McArthur, M. L. Sogin, and et al., *Evidence for lateral transfer of genes encoding ferredoxins, nitroreductases, nadh oxidase, and alcohol dehydrogenase 3 from anaerobic prokaryotes to giardia lamblia and entamoeba histolytica.*, Eukaryot. Cell **1** (2002), 181–190.

[283] Masaru K. Nobu, Takashi Narihiro, Ran Mei, Yoichi Kamagata, Patrick K. H. Lee, Po-Heng Lee, Michael J. McInerney, and Wen-Tso Liu, *Catabolism and interactions of uncultured organisms shaped by eco-thermodynamics in methanogenic bioprocesses*, Microbiome **8** (2020), no. 1, 111–111.

[284] H. Ochman, J. G. Lawrence, and E. A. Groisman, *Lateral gene transfer and the nature of bacterial innovation*, Nature **405** (2000), 299–304.

[285] M. Ogawa, S. Fujitani, X. Mao, S. Inouye, and T. Komano, *Frua, a putative transcription factor essential for the development of myxococcus xanthus*, Mol Microbiol **22** (1996), no. 4, 757–67.

[286] M. V. Omelchenko, Y. I. Wolf, E. K. Gaidamakova, V. Y. Matrosova, A. Vasilenko, M. Zhai, M. J. Daly, E. V. Koonin, and K. S. Makarova, *Comparative genomics of thermus thermophilus and deinococcus radiodurans: divergent routes of adaptation to thermophily and radiation resistance*, BMC Evol. Biol. **5** (2005), 57.

[287] Gregory S. Orf, Christopher Gisriel, and Kevin E. Redding, *Evolution of photosynthetic reaction centers: insights from the structure of the heliobacterial reaction center*, Photosynthesis Research **138** (2018), no. 1, 11–37.

[288] K. Ozawa, T. Meikari, K. Motohashi, M. Yoshida, and H. Akutsu, *Evidence for the presence of an f-type atp synthase involved in sulfate respiration in desulfovibrio vulgaris*, Journal of bacteriology **182** (2000), no. 8, 2200–2206.

[289] Mark J. Pallen and Brendan W. Wren, *Bacterial pathogenomics*, Nature **449** (2007), no. 7164, 835–842.

[290] L.W. Parfrey, D. J. G. Lahr, A.H. Knoll, and L. A. Katz, *Estimating the timing of early eukaryotic diversification with multigene molecular clocks*, Proc. Nat. Acad. Sci. USA **108** (2011), 13624– 13629.

[291] D. H. Parks and et al., *A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life.*, Nature Biotechnol. **36** (2018), 996–1004.

[292] D. H. Parks, M. Chuvochina, P. A. Chaumeil, C. Rinke, A. J. Mussig, and P. Hugenholtz, *A complete domain-to-species taxonomy for bacteria and archaea*, Nat Biotechnol (2020), NULL.

[293] D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P. A. Chaumeil, and P. Hugenholtz, *A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life*, Nat Biotechnol (2018), NULL.

[294] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, *Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes*, Genome Res **25** (2015), no. 7, 1043–55.

[295] D. H. Parks, C. Rinke, M. Chuvochina, P. A. Chaumeil, B. J. Woodcroft, P. N. Evans, P. Hugenholtz, and G. W. Tyson, *Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life*, Nat Microbiol **2** (2017), no. 11, 1533–1542.

[296] ———, *Author correction: Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life*, Nat Microbiol **3** (2018), no. 2, 253.

[297] D.H. Parks, M. Imelfort, C.T. Skennerton, P. Hugenholtz, and G.W. Tyson, *Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes,*, Genome Res. **25** (2015), 1043–1055.

[298] Donovan H. Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz, and Gene W. Tyson, *Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life*, Nature Microbiology **2** (2017), no. 11, 1533–1542.

[299] Sarah E. Partovi, Florence Mus, Andrew E. Gutknecht, Hunter A. Martinez, Brian P. Tripet, Bernd Markus Lange, Jennifer L. DuBois, and John W. Peters, *Coenzyme m biosynthesis in bacteria involves phosphate elimination by a functionally distinct member of the aspartase/fumarase superfamily*, Journal of Biological Chemistry **293** (2018), no. 14, 5236–5246.

[300] Natacha Pasche, Martin Schmid, Francisco Vazquez, Carsten J. Schubert, Alfred Wüest, John D. Kessler, Mary A. Pack, William S. Reeburgh, and Helmut Bürgmann, *Methane sources and sinks in lake kivu*, Journal of Geophysical Research: Biogeosciences **116** (2011), no. G3, NULL.

[301] Sandip Paul, Sumit K. Bag, Sabyasachi Das, Eric T. Harvill, and Chitra Dutta, *Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes*, Genome biology **9** (2008), no. 4, R70–R70.

[302] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay, *Scikit-learn: Machine learning in python*, J. Machine Learning Res. **12** (2011), 2825–2830.

[303] Maria Teresa Pellicer, Maria Felisa Nuñez, Juan Aguilar, Josefa Badia, and Laura Baldoma, *Role of 2-phosphoglycolate phosphatase of escherichia coli in metabolism of the 2-phosphoglycolate formed in dna repair*, Journal of Bacteriology **185** (2003), no. 19, 5815–5821.

[304] T. N. Petersen, S. Brunak, G. von Heijne, and H. Nielsen, *Signalp 4.0: discriminating signal peptides from transmembrane regions.*, Nat. Methods **8** (2011), 785–786.

[305] H. Philippe and C. J. Douady, *Horizontal gene transfer and phylogenetics*, Curr. Opin. Microbiol. **6** (2003), 498–505.

[306] B. Pivato, S. Mazurier, P. Lemanceau, S. Siblot, G. Berta, C. Mougel, and D. van Tuinen, *Medicago species affect the community composition of arbuscular mycorrhizal fungi associated with roots*, New Phytol **176** (2007), no. 1, 197–210.

[307] Arjan Pol, Klaas Heijmans, Harry R. Harhangi, Dario Tedesco, Mike S. M. Jetten, and Huub J. M. Op den Camp, *Methanotrophy below ph 1 by a new verrucomicrobia species*, Nature **450** (2007), no. 7171, 874–878.

[308] A. Prakash, M. Jeffryes, A. Bateman, and R. D. Finn, *The hmmer web server for protein sequence similarity search*, Curr Protoc Bioinformatics **60** (2017), 3.15.1–3.15.23.

[309] M. N. Price, P. S. Dehal, and A. P. Arkin, *Fasttree 2–approximately maximum-likelihood trees for large alignments*, PLoS One **5** (2010), no. 3, e9490.

[310] Roger C. Prince, Tivkaa J. Amande, and Terry J. McGenity, *Prokaryotic hydrocarbon degraders*, pp. 1–39, Springer International Publishing, Cham, 2019.

[311] P. Puigbo, A. Pasamontes, and S. Garcia-Vallve, *Gaining and losing the thermophilic adaptation in prokaryotes*, Trends Genet. **24** (2008), 10–14.

[312] Sonia Purin and Matthias C. Rillig, *The arbuscular mycorrhizal fungal protein glomalin: Limitations, progress, and a new hypothesis for its function*, Pedobiologia **51** (2007), no. 2, 123–130.

[313] Alejandro Pérez-de Luque, Stefanie Tille, Irene Johnson, David Pascual-Pardo, Jurriaan Ton, and Duncan D. Cameron, *The interactive effects of arbuscular mycorrhiza and plant growth-promoting rhizobacteria synergistically enhance host plant defences against pathogens*, Scientific Reports **7** (2017), no. 1, 16409.

[314] Q. Qian and P. J. Keeling, *Diplonemid glyceraldehyde–3-phosphate dehydrogenase (gapdh) and prokaryoteto- eukaryote lateral gene transfer.*, Protist **152** (2001), 193–201.

[315] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner, *The silva ribosomal rna gene database project: improved data processing and web-based tools*, Nucleic Acids Res **41** (2013), no. Database issue, D590–6.

[316] J. R. Quayle and N. Pfennig, *Utilization of methanol by rhodospirillaceae*, Arch Microbiol **102** (1975), no. 3, 193–8.

[317] S. Ranjan, Z. R. Todd, J. D. Sutherland, and D. D. Sasselov, *Sulfidic anion concentrations on early earth for surficial origins-of-life chemistry*, Astrobiology **18** (2018), 1023–1041.

[318] N. D. Rawlings, A. J. Barrett, P. D. Thomas, X. Huang, A. Bateman, and R. D. Finn, *The merops database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the panther database*, Nucleic Acids Res **46** (2018), no. D1, D624–D632.

[319] G. Recorbet, H. Rogniaux, V. Gianinazzi-Pearson, and E. DumasGaudot, *Fungal proteins in the extra-radical phase of arbuscular mycorrhiza: a shotgun proteomic picture*, New Phytol **181** (2009), no. 2, 248–260.

[320] Dirk Redecker, Arthur Schüßler, Herbert Stockinger, Sidney L. Stürmer, Joseph B. Morton, and Christopher Walker, *An evidence-based consensus for the classification of arbuscular mycorrhizal fungi (glomeromycota)*, Mycorrhiza **23** (2013), no. 7, 515–531.

[321] H. Reichenbach, *The ecology of the myxobacteria*, Environ Microbiol **1** (1999), no. 1, 15–21.

[322] G. Ricard, N. R. McEwan, B. E Dutilh, J-P. Jouany, D. Macheboeuf, M. Mitsumori, F. M. McIntosh, T. Michalowski, T. Nagamine, N. Nelson, C. J. Newbold, E. Nsabimana, A.Takenaka, N. A. Thomas, K. Ushida, J/ HP Hackstein, and M. A. Huynen, *Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment*, BMC Genomics **7** (2006), 22.

[323] T. A. Richards and A. Monier, *A tale of two tradigrades*, Proc. Nat. Acad. Sci. USA **113** (2017), 4892–4894.

[324] A. O. Richardson and J. D. Palmer, *Horizontal gene transfer in plants.*, J. Exp. Bot. **58** (2007), 1–9.

[325] P. Ricke, C. Erkel, M. Kube, R. Reinhardt, and W. Liesack, *Comparative analysis of the conventional and novel pmo (particulate methane monooxygenase) operons from methylocystis strain sc2*, Appl Environ Microbiol **70** (2004), no. 5, 3055–63.

[326] Matthias C. Rillig, Bruce A. Caldwell, Han A. B. Wösten, and Philip Sollins, *Role of proteins in soil carbon and nitrogen storage: controls on persistence*, Biogeochemistry **85** (2007), no. 1, 25–44.

[327] C. Rinke, P. Schwientek, A. Sczyrba, N. N. Ivanova, I. J. Anderson, J. F. Cheng, A. Darling, S. Malfatti, B. K. Swan, E. A. Gies, J. A. Dodsworth, B. P. Hedlund, G. Tsiamis, S. M. Sievert, W. T. Liu, J. A. Eisen, S. J. Hallam, N. C. Kyrpides, R. Stepanauskas, E. M. Rubin, P. Hugenholtz, and T. Woyke, *Insights into the phylogeny and coding potential of microbial dark matter*, Nature **499** (2013), no. 7459, 431–7.

[328] R. Rinke, F. Rubino, L. F Messer, N. Youssef, D. H. Parks, M. Chuvochina, M. Brown, T. Jeffries, G. W Tyson, J. R Seymour, and P. Hugenholtz, *A phylogenomic and ecological analysis of the globally abundant marine group ii archaea (ca. poseidoniales ord. nov.)*, The ISME J. **13** (2019), 663–675.

[329] S. J. Robbins, W. Song, J. P. Engelberts, B. Glasl, B. M. Slaby, J. Boyd, E. Marangon, E. S. Botté, P. Laffy, T. Thomas, and N. S. Webster, *A genomic view of the microbiome of coral reef demosponges*, The ISME Journal **15** (2021), no. 6, 1641–1654.

[330] Xavier Robert and Patrice Gouet, *Deciphering key features in protein structures with the new endscript server*, Nucleic Acids Research **42** (2014), no. W1, W320–W324.

[331] M. Robinson, B. Son, D. Kroos, and L. Kroos, *Transcription factor mrpc binds to promoter regions of hundreds of developmentally-regulated genes in myxococcus xanthus*, BMC Genomics **15** (2014), 1123.

[332] Fauziah F. Rochman, Miye Kwon, Roshan Khadka, Ivica Tamas, Azriel Abraham Lopez-Jauregui, Andriy Sheremet, Angela V. Smirnova, Rex R. Malmstrom, Sukhwan Yoon, Tanja Woyke, Peter F. Dunfield, and Tobin J. Verbeke, *Novel copper-containing membrane monooxygenases (cummos) encoded by alkane-utilizing betaproteobacteria*, The ISME Journal **14** (2020), no. 3, 714–726.

[333] Diana M. Rooke and R. C. Shattock, *Effect of chloramphenicol and streptomycin on developmental stages of phytophthom infestans*, Microbiology **129** (1983), no. 11, 3401–3410.

[334] S. Rosendahl, P. McGee, and J. B. Morton, *Lack of global population genetic differentiation in the arbuscular mycorrhizal fungus glomus mosseae suggests a recent range expansion which may have coincided with the spread of agriculture*, Mol Ecol **18** (2009), no. 20, 4316–29.

[335] Carl L. Rosier, Andrew T. Hoye, and Matthias C. Rillig, *Glomalin-related soil protein: Assessment of current detection and quantification tools*, Soil Biology and Biochemistry **38** (2006), no. 8, 2205–2211.

[336] M. O. Ross, F. MacMillan, J. Wang, A. Nisthal, T. J. Lawton, B. D. Olafson, S. L. Mayo, A. C. Rosenzweig, and B. M. Hoffman, *Particulate methane monooxygenase contains only mononuclear copper centers*, Science **364** (2019), no. 6440, 566–570.

[337] Marta Royo-Llonch, Pablo Sánchez, Clara Ruiz-González, Guillem Salazar, Carlos Pedrós-Alió, Karine Labadie, Lucas Paoli, Samuel Chaffron, Damien Eveillard, Eric Karsenti, Shinichi Sunagawa, Patrick Wincker, Lee Karp-Boss, Chris Bowler, and Silvia G Acinas, *Ecogenomics of key prokaryotes in the arctic ocean*, bioRxiv (2020), 2020.06.19.156794.

[338] Maxim Rubin-Blum, Chakkiath Paul Antony, Lizbeth Sayavedra, Clara Martínez-Pérez, Daniel Birgel, Jörn Peckmann, Yu-Chen Wu, Paco Cardenas, Ian MacDonald, Yann Marcon, Heiko Sahling, Ute Hentschel, and Nicole Dubilier, *Fueled by methane: deep-sea sponges from asphalt seeps gain their nutrition from methane-oxidizing symbionts*, The ISME Journal **13** (2019), no. 5, 1209–1225.

[339] G. Rufyikiri, S. Declerck, J. E. Dufey, and B. Delvaux, *Arbuscular mycorrhizal fungi might alleviate aluminium toxicity in banana plants*, New Phytologist **148** (2000), no. 2, 343–352.

[340] J. B. Russell, *Glucose toxicity in prevotella ruminicola: methylglyoxal accumulation and its effect on membrane physiology*, Applied and Environmental Microbiology **59** (1993), no. 9, 2844–2850.

[341] S. Sadekar, J. Raymond, and R. E. Blankenship, *Conservation of distantly related membrane proteins: photosynthetic reaction centers share a common structural core*, Mol Biol Evol **23** (2006), no. 11, 2001–7.

[342] Govind Prasad Sah and Daniel Wall, *Kin recognition and outer membrane exchange (ome) in myxobacteria*, Current Opinion in Microbiology **56** (2020), 81–88.

[343] Alessandra Salvioli, Stefano Ghignone, Mara Novero, Lorella Navazio, Francesco Venice, Paolo Bagnaresi, and Paola Bonfante, *Symbiosis with an endobacterium increases the fitness of a mycorrhizal fungus, raising its bioenergetic potential*, The Isme Journal **10** (2015), 130.

[344] R. A. Sanford, J. R. Cole, and J. M. Tiedje, *Characterization and description of anaeromyxobacter dehalogenans gen. nov., sp. nov., an aryl-halorespiring facultative anaerobic myxobacterium*, Appl Environ Microbiol **68** (2002), no. 2, 893–900.

[345] K. Schuchmann and V. Müller, *Energetics and application of heterotrophy in acetogenic bacteria*, Appl Environ Microbiol **82** (2016), no. 14, 4056–4069.

[346] G Schönknecht, A. P. Weber, and M. J. Lercher, *Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution*, Bioassays **36** (2013), 9–20.

[347] J. M. Senko, B. S. Campbell, J. R. Henricksen, M. S. Elshahed, T. A. Dewers, and L. R. Krumholz, *Barite deposition mediated by photootrophic sulfide-oxidizing bacteria*, Geochim. Cosmochim. Acta **68** (2004), 773–780.

[348] C. S. Sheik, S. Jain, and G. J. Dick, *Metabolic flexibility of enigmatic sar324 revealed through metagenomics and metatranscriptomics*, Environ Microbiol **16** (2014), no. 1, 304–17.

[349] A. Shevchenko, H. Tomas, J. Havlis, J. V. Olsen, and M. Mann, *In-gel digestion for mass spectrometric characterization of proteins and proteomes*, Nat Protoc **1** (2006), no. 6, 2856–60.

[350] L. Shimkets and C. R. Woese, *A phylogenetic analysis of the myxobacteria: basis for their classification*, Proceedings of the National Academy of Sciences **89** (1992), no. 20, 9459.

[351] Lawrence J. Shimkets, Martin Dworkin, and Hans Reichenbach, *The myxobacteria*, pp. 31–115, Springer New York, New York, NY, 2006.

[352] N. Shterzer and I. Mizrahi, *The animal gut as a melting pot for horizontal gene transfer*, Can J Microbiol **61** (2015), no. 9, 603–5.

[353] C. M. K. Sieber, A. J. Probst, A. Sharrar, B. C. Thomas, M. Hess, S. G. Tringe, and J. F. Banfield, *Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy*, Nat Microbiol **3** (2018), no. 7, 836–843.

[354] F. Sievers and D. G. Higgins, *Clustal omega for making accurate alignments of many protein sequences*, Protein Sci **27** (2018), no. 1, 135–145.

[355] J. Simon and P. M. H.Kroneck, *Microbial sulfite respiration*, Adv. Microbial Physiol. **62** (2013), 45–117.

[356] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, *Busco: assessing genome assembly and annotation completeness with single-copy orthologs*, Bioinformatics **31** (2015), 3210–3212.

[357] S. M. Smith, S. Rawat, J. Telser, B. M. Hoffman, T. L. Stemmler, and A. C. Rosenzweig, *Crystal structure and characterization of particulate methane monooxygenase from methylocystis species strain m*, Biochemistry **50** (2011), no. 47, 10231–40.

[358] Sally E. Smith and F. Andrew Smith, *Roles of arbuscular mycorrhizas in plant nutrition and growth: New paradigms from cellular to ecosystem scales*, Annual Review of Plant Biology **62** (2011), no. 1, 227–250.

[359] Sally E. Smith, F. Andrew Smith, and Iver Jakobsen, *Mycorrhizal fungi can dominate phosphate supply to plants irrespective of growth responses*, Plant Physiology **133** (2003), no. 1, 16.

[360] _____, *Functional diversity in arbuscular mycorrhizal (am) symbioses: the contribution of the mycorrhizal p uptake pathway is not correlated with mycorrhizal responses in growth or total p uptake*, New Phytologist **162** (2004), no. 2, 511–524.

[361] Y. Song, D. Chen, K. Lu, Z. Sun, and R. Zeng, *Enhanced tomato disease resistance primed by arbuscular mycorrhizal fungus*, Front Plant Sci **6** (2015), 786.

[362] D. Y. Sorokin, T. P. Tourova, M. Mussmann, and G. Muyzer, *Dethiobacter alkaliphilus gen. nov. sp. nov., and desulfurivibrio alkaliphilus gen. nov. sp. nov.: two novel representatives of reductive sulfur cycle from soda lakes*, Extremophiles **12** (2008), no. 3, 431–9.

[363] S. M. Soucy, J. Huang, and J. P. Gogarten, *Horizontal gene transfer: building the web of life*, Nat Rev Genet **16** (2015), no. 8, 472–82.

[364] A. M. Spain, M. S. Elshahed, F. Z. Najar, and L. R. Krumholz, *Metatranscriptomic analysis of a high-sulfide aquatic spring reveals insights into sulfur cycling and unexpected aerobic metabolism*, PeerJ **3** (2015), e1259.

[365] A. M. Spain, F. Z. Najar, L. R. Krumholz, and M. S. Elshahed, *Comparative metratranscriptomic analysis of a high-sulfide spring reveals insight into sulfur cycling pathways and unexpected aerobic metabolism*, Under Review (2014), NULL.

[366] Cathrin Spröer, Hans Reichenbach, and Erko Stackebrandt, *The correlation between morphological and phylogenetic classification of myxobacteria*, International Journal of Systematic and Evolutionary Microbiology **49** (1999), no. 3, 1255–1262.

[367] M. St-Arnaud, C. Hamel, B. Vimard, M. Caron, and J. A. Fortin, *Enhanced hyphal growth and spore production of the arbuscular mycorrhizal fungus glomus intraradices in an in vitro system in the absence of host roots*, Mycological Research **100** (1996), no. 3, 328–332.

[368] Courtney W. Stairs, Laura Eme, Sergio A. Muñoz-Gómez, Alejandro Cohen, Graham Dellaire, Jennifer N. Shepherd, James P. Fawcett, and Andrew J. Roger, *Microbial eukaryotes have adapted to hypoxia by horizontal acquisitions of a gene involved in rhodoquinone biosynthesis*, eLife **7** (2018), e34292.

[369] Alexandros Stamatakis, *Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies*, Bioinformatics **30** (2014), no. 9, 1312–1313.

[370] Robert D. Stewart, Marc D. Auffret, Amanda Warr, Andrew H. Wiser, Maximilian O. Press, Kyle W. Langford, Ivan Liachko, Timothy J. Snelling, Richard J. Dewhurst, Alan W. Walker, Rainer Roehe, and Mick Watson, *Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen*, Nature Communications **9** (2018), no. 1, 870.

[371] M. Stolzer, H. Lai, M. Xu, D. Sathaye, B. Vernot, and D. Durand, *Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees*, Bioinformatics **28** (2012), no. 18, i409–i415.

[372] G. L. Sun, Z. F. Yang, A. Ishwar, and J. L. Huang, *Algal genes in the closest relatives of animals*, Mol. Biol. Evol. **27** (2010), 2879–2889.

[373] X. Sun, W. Chen, S. Ivanov, A. M. MacLean, H. Wight, T. Ramaraj, J. Mudge, M. J. Harrison, and Z. Fei, *Genome and evolution of the arbuscular mycorrhizal fungus diversispora epigaea (formerly glomus versiforme) and its bacterial endosymbionts*, New Phytol **221** (2019), no. 3, 1556–1573.

[374] Jan H Swiegers, Frédéric M Vaz, Isak S Pretorius, Ronald J.A Wanders, and Florian F Bauer, *Carnitine biosynthesis in neurospora crassa: identification of a cdna coding for e-n-trimethyllysine hydroxylase and its functional expression in saccharomyces cerevisiae*, FEMS Microbiology Letters **210** (2002), no. 1, 19–23.

[375] M. Syvanen, *Evolutionary implications of horizontal gene transfer*, Annu. Rev. Genet. **46** (2012), 341–358.

[376] D. Søndergaard, C. N. Pedersen, and C. Greening, *Hyddb: A web tool for hydrogenase classification and analysis*, Sci Rep **6** (2016), 34212.

[377] C. D. Taylor and R. S. Wolfe, *Structure and methylation of coenzyme m(hsch2ch2so3)*, J Biol Chem **249** (1974), no. 15, 4879–85.

[378] J. W. Taylor and M. L. Berbee, *Dating divergences in the fungal tree of life: review and new analyses*, Mycologia **98** (2006), no. 6, 838–49.

[379] Susanne Thiery and Christine Kaimer, *The predation strategy of myxococcus xanthus*, Frontiers in microbiology **11** (2020), 2–2.

[380] Sara H. Thomas, Ryan D. Wagner, Adrian K. Arakaki, Jeffrey Skolnick, John R. Kirby, Lawrence J. Shimkets, Robert A. Sanford, and Frank E. Löffler, *The mosaic genome of anaeromyxobacter dehalogenans strain 2cp-c suggests an aerobic common ancestor to the delta-proteobacteria*, PloS one **3** (2008), no. 5, e2103–e2103.

[381] Fei Tian, Yong Yu, Bo Chen, Huirong Li, Yu-Feng Yao, and Xiao-Kui Guo, *Bacterial, archaeal and eukaryotic diversity in arctic sediment as revealed by 16s rrna and 18s rrna gene clone libraries analysis*, Polar Biology **32** (2009), no. 1, 93–103.

[382] R. M. Tian, W. Zhang, L. Cai, Y. H. Wong, W. Ding, and P. Y. Qian, *Genome reduction and microbe-host interactions drive adaptation of a sulfur-oxidizing bacterium associated with a cold seep sponge*, mSystems **2** (2017), no. 2, NULL.

[383] E. Tisserant, A. Kohler, P. Dozolme-Seddas, R. Balestrini, K. Benabdellah, A. Colard, D. Croll, C. Da Silva, S. K. Gomez, R. Koul, N. Ferrol, V. Fiorilli, D. Formey, P. Franken, N. Helber, M. Hijri, L. Lanfranco, E. Lindquist, Y. Liu, M. Malbreil, E. Morin, J. Poulain, H. Shapiro, D. van Tuinen, A. Waschke, C. Azcon-Aguilar, G. Becard, P. Bonfante, M. J. Harrison, H. Kuster, P. Lammers, U. Paszkowski, N. Requena, S. A. Rensing, C. Roux, I. R. Sanders, Y. Shachar-Hill, G. Tuskan, J. P. Young, V. Gianinazzi-Pearson, and F. Martin, *The transcriptome of the arbuscular mycorrhizal fungus glomus intraradices (daom 197198) reveals functional tradeoffs in an obligate symbiont*, New Phytol **193** (2012), no. 3, 755–69.

[384] E. Tisserant, M. Malbreil, A. Kuo, A. Kohler, A. Symeonidi, R. Balestrini, P. Charron, N. Duensing, N. Frei dit Frey, V. Gianinazzi-Pearson, L. B. Gilbert, Y. Handa, J. R. Herr, M. Hijri, R. Koul, M. Kawaguchi, F. Krajinski, P. J. Lammers, F. G. Masclaux, C. Murat, E. Morin, S. Ndikumana, M. Pagni, D. Petitpierre, N. Requena, P. Rosikiewicz, R. Riley, K. Saito, H. San Clemente, H. Shapiro, D. van Tuinen, G. Becard, P. Bonfante, U. Paszkowski, Y. Y. Shachar-Hill, G. A. Tuskan, J. P. Young, I. R. Sanders, B. Henrissat, S. A. Rensing, I. V. Grigoriev, N. Corradi, C. Roux, and F. Martin, *Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis*, Proc Natl Acad Sci U S A **110** (2013), no. 50, 20117–22.

[385] JM Tsuji, NA Shaw, S Nagashima, JJ Venkiteswaran, SL Schiff, S Hanada, M Tank, and JD Neufeld, *Anoxygenic phototrophic ¡em¿chloroflexota¡/em¿ member uses a type i reaction center*, bioRxiv (2020), 2020.07.07.190934.

[386] J. B. van Beilen and E. G. Funhoff, *Alkane hydroxylases involved in microbial alkane degradation*, Appl Microbiol Biotechnol **74** (2007), no. 1, 13–21.

[387] C. Vannini, A. Carpentieri, A. Salvioli, M. Novero, M. Marsoni, L. Testa, M. C. de Pinto, A. Amoresano, F. Ortolani, M. Bracale, and P. Bonfante, *An interdomain network: the endobacterium of a mycorrhizal fungus promotes antioxidative responses in both fungal and plant hosts*, New Phytol **211** (2016), no. 1, 265–75.

[388] Erik Verbruggen, Marcel G. A. Van Der Heijden, James T. Weedon, George A. Kowalchuk, and Wilfred F. M. RÖLing, *Community assembly, species richness and nestedness of arbuscular mycorrhizal fungi in agricultural soils*, Molecular Ecology **21** (2012), no. 10, 2341–2353.

[389] S. Voruganti, J. T. Kline, M. J. Balch, J. Rogers, R. L. Matts, and S. D. Hartson, *Proteomic profiling of hsp90 inhibitors*, Methods Mol Biol **1709** (2018), 139–162.

[390] D. W. Waite, M. Chuvochina, C. Pelikan, D. H. Parks, P. Yilmaz, M. Wagner, A. Loy, T. Naganuma, R. Nakai, W. B. Whitman, M. W. Hahn, J. Kuever, and P. Hugenholtz, *Proposal to reclassify the proteobacterial classes deltaproteobacteria and oligoflexia, and the phylum thermodesulfobacteria into four phyla reflecting major functional capabilities*, Int J Syst Evol Microbiol **70** (2020), no. 11, 5972–6016.

[391] David W. Waite, Inka Vanwonterghem, Christian Rinke, Donovan H. Parks, Ying Zhang, Ken Takai, Stefan M. Sievert, Jörg Simon, Barbara J. Campbell, Thomas E. Hanson, Tanja Woyke, Martin G. Klotz, and Philip Hugenholtz, *Comparative genomic analysis of the class epsilonproteobacteria and proposed reclassification to epsilonbacteraeota (phyl. nov.)*, Frontiers in Microbiology **8** (2017), no. 682, NULL.

[392] Wanpeng Wang, Zhenyu Li, Lingyu Zeng, Chunming Dong, and Zongze Shao, *The oxidation of hydrocarbons by diverse heterotrophic and mixotrophic bacteria that inhabit deep-sea hydrothermal ecosystems*, The ISME Journal **14** (2020), no. 8, 1994–2006.

[393] X Wang, X Liu, and J. Z. Groenewald, *Phylogeny of anaerobic fungi (phylum neocallimastigomycota), with contributions from yak in china*, Antonie Van Leeuwenhoek **110** (2017), 87–103.

[394] Zhengping Wang, Dong Zeng, and William H. Patrick, *Methane emissions from natural wetlands*, Environmental Monitoring and Assessment **42** (1996), no. 1, 143–161.

[395] Zhi-Gang Wang, Yin-Li Bi, Bin Jiang, Yryszhan Zhakypbek, Su-Ping Peng, Wen-Wen Liu, and Hao Liu, *Arbuscular mycorrhizal fungi enhance soil carbon sequestration in the coalfields, northwest china*, Scientific Reports **6** (2016), 34336.

[396] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. de Beer, C. Rempfer, L. Bordoli, R. Lepore, and T. Schwede, *Swiss-model: homology modelling of protein structures and complexes*, Nucleic Acids Res **46** (2018), no. W1, W296–w303.

[397] A. M. Waterhouse, J. B. Procter, D. M. Martin, M. Clamp, and G. J. Barton, *Jalview version 2–a multiple sequence alignment editor and analysis workbench*, Bioinformatics **25** (2009), no. 9, 1189–91.

[398] D. E. Whitworth and A. Zwarycz, *A genomic survey of signalling in the myxococcaceae*, Microorganisms **8** (2020), no. 11, NULL.

[399] J. Wiedenbeck and F. M. Cohan, *Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches*, FEMS Microbiol. Rev. **35** (2011), 957–976.

[400] G. W. Wilson, C. W. Rice, M. C. Rillig, A. Springer, and D. C. Hartnett, *Soil aggregation and carbon sequestration are tightly correlated with the abundance of arbuscular mycorrhizal fungi: results from long-term field experiments*, Ecol Lett **12** (2009), no. 5, 452–61.

[401] J. H. Wisecaver, M. L. Brosnahan, and J. D. Hackett, *Horizontal gene transfer is a significant driver of gene innovation in dinoflagellates*, Genome Biol. Evol. **5** (2013), 2368–2381.

[402] J. R. Wisniewski, A. Zougman, N. Nagaraj, and M. Mann, *Universal sample preparation method for proteome analysis*, Nat Methods **6** (2009), no. 5, 359–62.

[403] S. F. Wright and A. Upadhyaya, *Extraction of an abundant and unusual protein from soil and comparison with hyphal protein of arbhuscular mycorrhizal fungi*, Soil Science **161** (1996), 575–586.

[404] K. C. Wrighton, B. C. Thomas, I. Sharon, C. S. Miller, C. J. Castelle, N. C. VerBerkmoes, M. J. Wilkins, R. L. Hettich, M. S. Lipton, K. H. Williams, P. E. Long, and J. F. Banfield, *Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla*, Science **337** (2012), no. 6102, 1661–5.

[405] Y.W. Wu, B.A. Simmons, and S.W. Singer, *Maxbin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets,*, Bioinformatics **32** (2016), 605–607.

[406] T. Wubet, M. Weiss, I. Kottke, D. Teketay, and F. Oberwinkler, *Phylogenetic analysis of nuclear small subunit rdna sequences suggests that the endangered african pencil cedar, juniperus procera, is associated with distinct members of glomeraceae*, Mycol Res **110** (2006), no. Pt 9, 1059–69.

[407] A. Yadav, J. C. Borrelli, M. S. Elshahed, and N. H. Youssef, *Genomic analysis of family uba6911 (group 18 acidobacteria) expands the metabolic capacities of the phylum and highlights adaptations to terrestrial habitats*, Appl Environ Microbiol **87** (2021), no. 17, e0094721.

[408] E. Yamamoto, H. Muramatsu, and K. Nagai, *Vulgatibacter incomptus gen. nov., sp. nov. and labilithrix luteola gen. nov., sp. nov., two myxobacteria isolated from soil in yakushima island, and the description of vulgatibacteraceae fam. nov., labilitrichaceae fam. nov. and anaeromyxobacteraceae fam. nov*, Int J Syst Evol Microbiol **64** (2014), no. Pt 10, 3360–3368.

[409] Lin Ye, Ran Mei, Wen-Tso Liu, Hongqiang Ren, and Xu-Xiang Zhang, *Machine learning-aided analyses of thousands of draft genomes reveal specific features of activated sludge processes*, Microbiome **8** (2020), no. 1, 16.

[410] Yanbin Yin, Xizeng Mao, Jincai Yang, Xin Chen, Fenglou Mao, and Ying Xu, *dbcan: a web resource for automated carbohydrate-active enzyme annotation*, Nucleic Acids Research **40** (2012), no. Web Server issue, W445–W451.

[411] N. H. Youssef, P. C. Blainey, S. R. Quake, and M. S. Elshahed, *Partial genome assembly for a candidate division op11 single cell from an anoxic spring (zodletone spring, oklahoma)*, Appl Environ Microbiol **77** (2011), no. 21, 7804–14.

[412] N. H. Youssef, M. B. Couger, C. G. Struchtemeyer, A. S. Liggenstoffer, R. A. Prade, F. Z. Najar, H. K. Atiyeh, M. R. Wilkins, and M. S. Elshahed, *Genome of the anaerobic fungus orpinomyces sp. c1a reveals the unique evolutionary history of a remarkable plant biomass degrader*, Appl. Environ. Microbiol. **79** (2013), 4620–4634.

[413] _____, *The genome of the anaerobic fungus orpinomyces sp. strain c1a reveals the unique evolutionary history of a remarkable plant biomass degrader*, Appl Environ Microbiol **79** (2013), no. 15, 4620–34.

[414] N. H. Youssef, C. Rinke, R. Stepanauskas, I. Farag, T. Woyke, and M. S. Elshahed, *Insights into the metabolism, lifestyle and putative evolutionary history of the novel archaeal phylum 'diapherotrites'*, ISME J **9** (2015), 447–460.

[415] Noha H. Youssef, Ibrahim F. Farag, Sydney Rudy, Ace Mulliner, Kara Walker, Ford Caldwell, Malik Miller, Wouter Hoff, and Mostafa Elshahed, *The wood–ljungdahl pathway as a key component of metabolic versatility in candidate phylum bipolaricaulota (acetothermia, op1)*, Environmental Microbiology Reports **11** (2019), no. 4, 538–547.

[416] Jing Zhang, Fang Wang, Rongxiao Che, Ping Wang, Hanke Liu, Baoming Ji, and Xiaoyong Cui, *Precipitation shapes communities of arbuscular mycorrhizal fungi in tibetan alpine steppe*, Scientific Reports **6** (2016), 23488.

[417] H. Zheng and H. Wu, *Gene-centric association analysis for the correlation between the guanine-cytosine content levels and temperature range conditions of prokaryotic species*, BMC Bioinformatics **11 Suppl 11** (2010), no. Suppl 11, S7.

[418] Xiu-wen Zhou, Shu-guang Li, Wei Li, De-ming Jiang, Kui Han, Zhi-hong Wu, and Yue-zhong Li, *Myxobacterial community is a predominant and highly diverse bacterial group in soil niches*, Environmental Microbiology Reports **6** (2014), no. 1, 45–56.

[419] Zhichao Zhou, Patricia Q. Tran, Kristopher Kieft, and Karthik Anantharaman, *Genome diversification in globally distributed novel marine proteobacteria is linked to environmental adaptation*, The ISME Journal **14** (2020), no. 8, 2060–2077.

[420] David R. Zusman, Ansley E. Scott, Zhaomin Yang, and John R. Kirby, *Chemosensory pathways, motility and development in myxococcus xanthus*, Nature Reviews Microbiology **5** (2007), no. 11, 862–872.

VITA

Chelsea Louvoun Murphy

Candidate for the Degree of

Doctor of Philosophy

Dissertation: LEVERAGING -OMICS BASED APPROACHES TO EXPLORE ENVIRONMENTS: A LOOK AT TWO DOMAINS OF LIFE

Major Field: Microbiology, Cell and Molecular Biology

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Microbiology, Cell and Molecular Biology at Oklahoma State University, Stillwater, Oklahoma in May, 2022.

Completed the requirements for the Bachelor of Science in Microbiology, Cell and Molecular Biology at Oklahoma State University, Stillwater, Oklahoma in 2017.

Professional Memberships:

American Society of Microbiology